2 Data Selection

2.1 Data Source

The data contains the neibourhood's location data and the venue data. The location data is webscreping from wikipedia:https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, the location is from http://cocl.us/Geospatial_data. The venue data is request from the foursquare.

2.2 Data Selection

The data contains three levels. The first level is the location of each neighbourhood. The second level is the venues belong to each neighbourhood, and the third level is the properties of each venue.

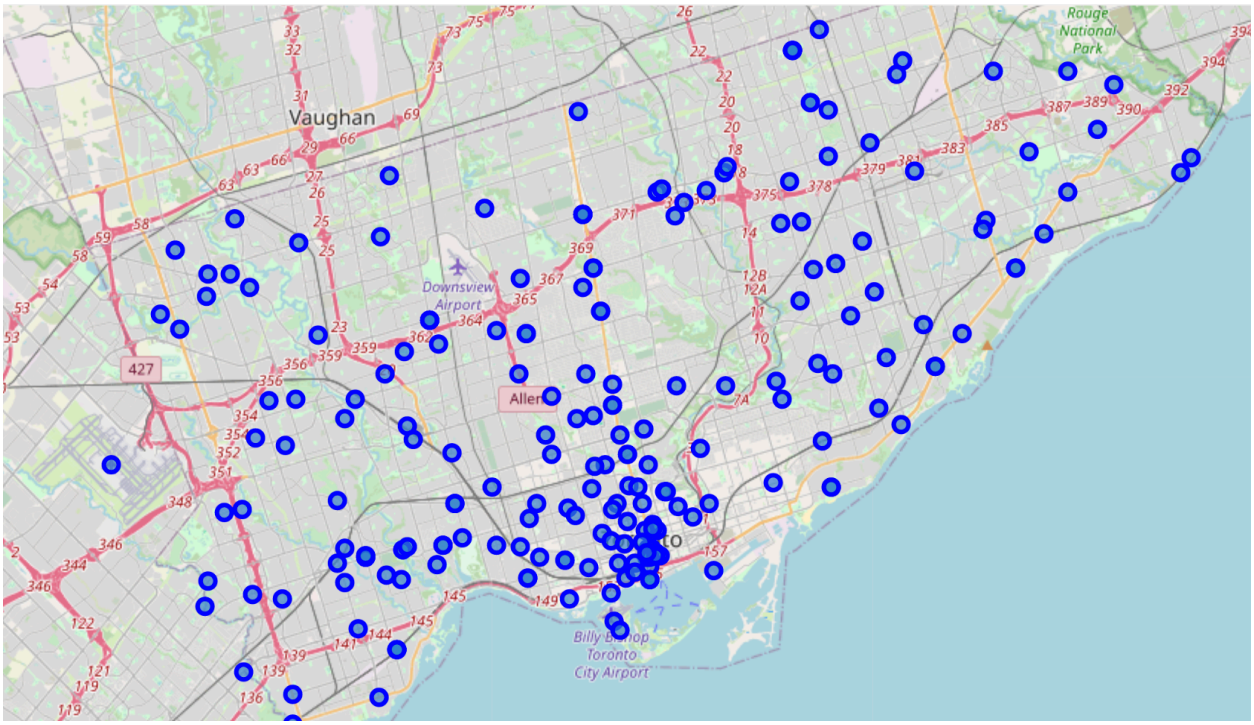
We will focusing on the venue which belong to the food catagory. If the time is allowed, we will separate the restaurant into detailed catagory to analyze. First, lets look at the first level data.

2.2.1 First level data

	Neighbourhood	Borough	Postal Code	Latitude	Longitude
0	Parkwoods	North York	M3A	43.758800	-79.320197
1	Victoria Village	North York	M4A	43.732658	-79.311189
2	Regent Park	Downtown Toronto	M5A	43.660706	-79.360457
3	Harbourfront	Downtown Toronto	M5A	43.640080	-79.380150
4	Lawrence Manor	North York	M6A	43.722079	-79.437507

Clean Data

There are 3 neighbourhood names are invalid for business, which are "Business reply mail Processing Centre", "South Central Letter Processing Plant Toronto" and "Canada Post Gateway Processing Centre". They need to be droped. The location of the neighbourhood need to be relocated since it's previously located by their postal code. The relocated neighbourhood location is shown below.



2.2.2 Second level data

The second level data includes the venues list in the Neighbourhood. Using the explore endpoint to request the food-category venues information from foursquare. The first 5 rows of the second level data is shown below. It has 864 rows.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	id
0	Parkwoods	43.758800	-79.320197	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant	4b8991cbf964a520814232e3
1	Parkwoods	43.758800	-79.320197	Pizza Pizza	43.760231	-79.325666	Pizza Place	4b5f7253f964a520d7ba29e3
2	Parkwoods	43.758800	-79.320197	Spicy Chicken House	43.760639	-79.325671	Chinese Restaurant	4c0150f4716bc9b65b9dbb55
3	Parkwoods	43.758800	-79.320197	Allwyn's	43.761000	-79.325478	Caribbean Restaurant	4c729f4aad69b60c81ee83b9
4	Victoria Village	43.732658	-79.311189	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant	4b8991cbf964a520814232e3

The second level data also contains a full venue list. It not only contains the restaurant list but also contains other venues such as gym and gas stations. The first 5 rows of the second level data is shown below. It has 2592 rows.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	id
0	Parkwoods	43.7588	-79.320197	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant	4b8991cbf964a520814232e3
1	Parkwoods	43.7588	-79.320197	LCBO	43.757774	-79.314257	Liquor Store	4bdccf4cafe8c9b6da285185
2	Parkwoods	43.7588	-79.320197	Shoppers Drug Mart	43.760857	-79.324961	Pharmacy	4c422e48e26920a1a4ad5fe7
3	Parkwoods	43.7588	-79.320197	Petro-Canada	43.757950	-79.315187	Gas Station	4c361e9118e72d7fca4714f5
4	Parkwoods	43.7588	-79.320197	Pizza Pizza	43.760231	-79.325666	Pizza Place	4b5f7253f964a520d7ba29e3

2.2.3 Third level data

The third level data comes from the foursquare API, by using premium request. Each line corresponds to the same line in second level restaurant data. The first 5 rows of the third level data is shown below. It has 864 rows.

	Tier	Rating	TipNumber	Likes	Catagory Number	Catagory
0	2	9.0	16	18	1	[Caribbean Restaurant]
1	1	6.3	2	2	1	[Pizza Place]
2	1	0.0	4	1	1	[Chinese Restaurant]
3	2	0.0	2	1	1	[Caribbean Restaurant]
4	2	9.0	16	18	1	[Caribbean Restaurant]

Clean the third level data

It is possible that we asked a venue for detailed properties but the response message lacks some data such as rating. We put zero for the lacking integer and float format data. The category column will always have non-void data.