Finding optimal location to open restaurant/grocery business

IBM Applied Data Science Capstone as a fulfilment for the IBM Data Science Professional Certificate Specialization

Sumudu Tennakoon

2018-01-02

Abstract

Neighborhoods in general possess characteristics are highly personalized to the demographics of inhabitants, environment, locality, etc. We can consider it as an entity with personality that can be used in a data driven location service. This project focus on conducting exploratory data analysis on a set of neighborhoods using the data collected from multiple different sources. Complements with a machine learning methods of clustering similar neighborhoods, this project attempts to find solutions to a client to locate places to establish their business.

Introduction

An international grocery and restaurant chain looking forward opening their business locations in the city of Toronto. They wanted to identify optimum locations having maximum businesses potential and required to generate business intelligence to form a strategy in establishing their new business locations. In the week 3 assignment we note that the Toronto city has neighborhoods each of them having different characteristics. Further examining the area along with other sources, it shows many of the location related services can provide by having a rich dataset to describe the neighborhoods.

This project will conduct analyzing population demographics, financial and household data in those neighborhoods and cluster them based on their similarity. It will also find the existing venues creating competition (e.g. Restaurants, Grocery stores) and other venues in the proximity which adds new businesses opportunities. The first half of the project was dedicated to extract data from different sources and build a dataset that can be used to solve the problem as well as can be applied to solve many other interesting data related problems in the city of Toronto.

Data

City of Toronto's Open Data Catalogue

This catalogue contains vast about of data in many different forms. The data is licensed under the Open Government License – Toronto. The following datasets were obtained and used in the project.

Table1: Datasets from City of Toronto's Open Data Catalogue

Name	Description	File Format

Population Demographics	Population data by neighborhood including population by age categories, culture and ethnicity.	csv
Safety	Crime incidents by neighborhood	xlsx->csv
Economy	Income, credit rating	xlsx->csv
Hosing	Households, House price	xlsx->csv

The data from Open Data Catalogue will be used to cluster neighborhoods based on their similarity characteristics. This will help the business to group neighborhoods when forming custom business strategies to their targeted neighborhoods. This data will also be used in finding the optimum business locations.

Geospatial Coordinates of Neighborhoods

To find the location centers for each neighborhood, a web scraping was carried out across number of Wikipedia pages starting from. The geolocations were extracted from each neighborhood's Wikipedia page and for few of them it had to locate manually referring to other map sources (Google Maps and Bing Maps).

Name	Description	File Format
Postal **Codes** Dataset	List of postal codes of Toronto, Canada web scrapped from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M	csv
Geospatial Coordinates of Postal Codes	Geographical coordinates of each postal code: http://cocl.us/Geospatial_data provided in the assignment.	csv

Foursquare APIs location data

The Foursquare API was used to extract information about venues using the explore option provided search key to query. Restaurant, grocery stores, entertainment (fun), hotels, shopping, parking was used as search keys to get a dataset of venues which are potentially connected to the client's biasness. This dataset was used to identify competitive business locations in each neighborhood (e.g. grocery stores, and restaurants) as well as venues which adds new businesses opportunities (e.g. Parking, Attractions, Shopping Malls, etc.). Figure 1 shows a map of neighborhood centers and venues of a selected neighborhood.

Name	Description	File Format
Venues	Location, Venue Name, Venue Type, Ratings (if any)	API

All source datasets were extracted, cleaned and prepared datasets to be used in the next phase of the project.

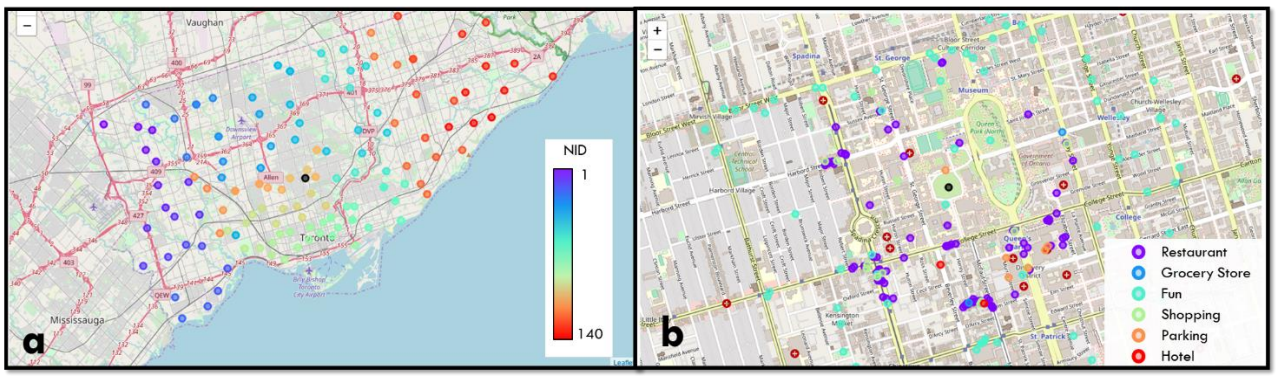

Figure 1: (a) Toronto neighborhoods (NID=Neighborhood ID), (b) Exploring neighborhood venues.

Methodology

Exploratory data analysis

Figure A1-A4 shows the results of exploratory data analysis from this dataset. The distributions of each variable give an idea of overall characteristics of the neighborhoods in Toronto. The box plot is used to have another view to the distribution and to get an idea of outliers presence in the dataset. Identification of top 10% (14/140) neighborhoods for each variable is also a useful outcome for the client to form their businesses strategies.

Software Platform & tools

- IBM Watson Studio
- Python
- Jupiter Notebook
- Foursquare API (https://developer.foursquare.com)
- Folium Leaflet maps (https://github.com/python-visualization/folium)
- BeautifulSoup (https://www.crummy.com/software/BeautifulSoup/bs4)

Neighborhood clustering

Group similar neighborhoods based on their will help the client to have better understanding on similar neighborhoods that they can form limited number of business strategies and models. That also help if they want to apply same strategy or model to open the businesses location on multiple neighborhoods or to move existing one to a different location. An unsupervised machine learning method K-Means clustering has been utilized in achieving this.

Ranking Neighborhoods based on their characteristics

Since a repose variable is not available, we cannot use machine learning method to score and rank the neighborhoods. Therefore, characteristics which will positively and negatively has been identified among all the data fields. Then the neighborhoods were ranked by individual fields. To calculate the overall rank presented in this report, the following formula is used *a *[Sum of ranks of Positive Variables] – b *[Sum of rank of negative variables]*.

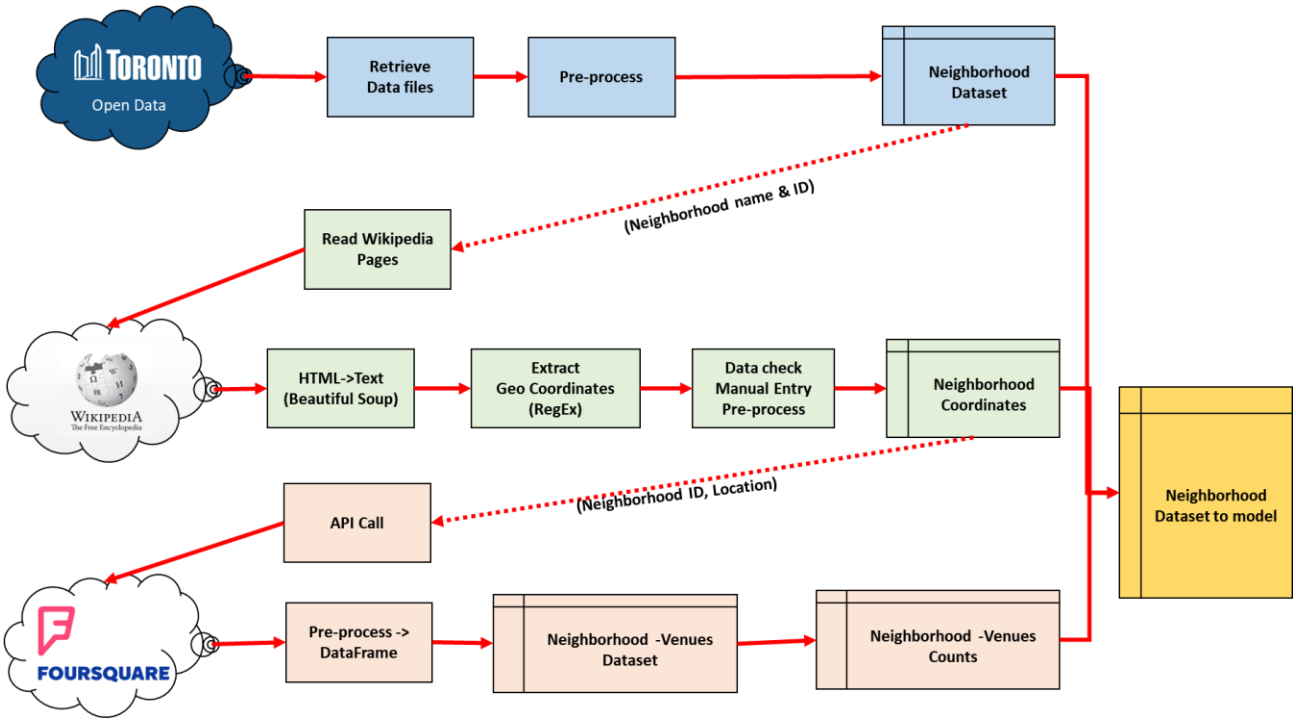

Figure 2: Process of data extraction and pre-processing.

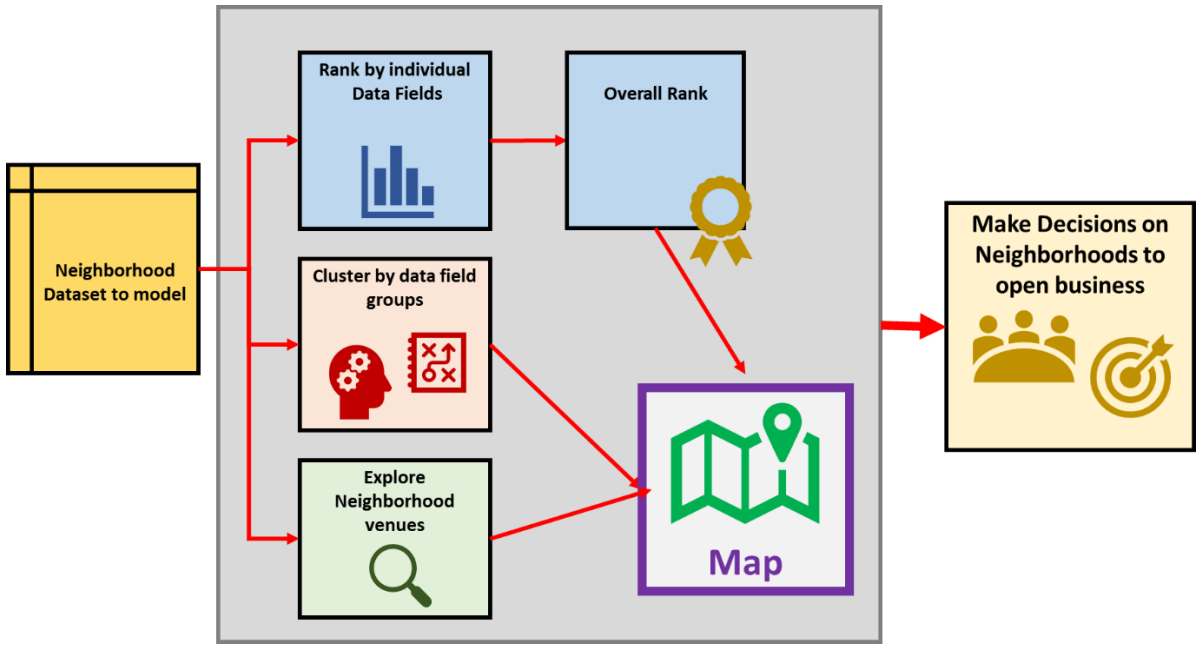

Figure 3: Process of data analysis and decision making from project results.

Results

Clustering is done based on different factors including Economy, Population, Locality, and Safety (Figure 4 and Figure 5). The dataset is created that client can make their own clustering using by combining different fields.

			Cluster				Rank													
Neighbourhood Id	Neighbourhood	TargetPopulation	Population	Economy	Locality	Safty	Overall Rank	Population 2016	Business target population	Diversity	Supportive venues	Average after tax household income 2015	Average household size	Debt Risk Score	Employment rate	Local Employment	Businesses	Competitive Venues	Home Prices	Break Enters Robberies Thefts	Total Major Crime Incidents
52	Bayview Village	15,355					1	50	41	15	40	47	109	25	71	37	51	69	52	115	126
53	Henry Farm	11,765					1	79	68	10	60	102	75	81	71	13	26	118	109	106	124
55	Thorncliffe Park	13,140					3	53	56	14	91	106	15	98	137	36	52	97	124	128	84
82	Niagara	27,620					4	15	9	95	7	42	137	62	1	21	35	10	100	102	35
42	Banbury-Don Mills	17,095					5	23	32	54	24	29	108	13	97	9	24	56	33	55	77
117	L'Amoreaux	28,870					6	5	7	16	84	25	18	38	129	63	54	56	113	19	48
59	Danforth East York	11,515					7	65	70	108	57	57	83	32	47	74	107	105	58	125	117
132	Malvern	30,020					8	7	6	36	136	24	5	98	98	34	32	100	135	11	14
87	High Park-Swansea	16,565					9	35	38	99	24	30	115	23	9	47	53	29	32	93	90
127	Bendale	20,190					10	18	20	44	70	60	37	77	117	16	21	75	120	25	24
118	Tam O'Shanter-Sullivan	17,410					11	25	30	27	63	31	42	43	126	72	73	58	96	52	71
137	Woburn	35,850					12	2	3	46	57	15	21	100	123	19	15	47	122	10	9
131	Rouge	31,900					13	4	4	47	63	9	1	77	67	23	20	49	89	15	25
37	Willowdale West	11,780					14	67	67	22	113	73	109	28	121	44	67	93	45	128	113
11	Eringate-Centennial-West Deane	12,170					15	58	63	85	101	49	37	16	69	79	93	113	90	67	99
130	Milliken	17,790					16	29	28	2	63	26	2	23	133	17	8	18	103	24	60
77	Waterfront Communities-The Island	57,625					17	1	1	79	15	12	138	58	2	2	2	21	93	8	1
116	Steeles	15,700					17	34	40	1	68	44	9	8	140	50	92	31	86	60	98
1	West Humber-Clairville	23,280					19	12	14	23	53	37	6	107	73	3	4	51	121	1	2
129	Agincourt North	19,240					20	20	23	3	48	36	8	29	132	43	69	14	105	30	69
126	Dorset Park	16,970					21	33	34	51	53	86	29	94	78	26	17	47	132	36	34
48	Hillcrest Village	10,495					22	68	84	6	97	68	51	16	134	30	56	63	74	86	107
104	Mount Pleasant West	22,650					23	19	15	87	63	34	136	50	6	24	14	35	57	47	50
38	Lansing-Westgate	11,605					23	74	69	55	32	46	80	38	46	28	44	26	30	114	103
128	Agincourt South-Malvern West	16,590					25	37	37	7	48	72	20	43	114	14	11	14	118	33	43
51	Willowdale East	38,250					26	3	2	4	32	19	103	36	92	20	22	27	46	17	27
17	Mimico (includes Humber Bay Shores)	25,325					27	11	11	93	97	18	130	71	20	33	33	59	88	28	30
105	Lawrence Park North	9,375					28	87	95	133	118	32	42	4	22	76	61	133	11	96	121
46	Pleasant View	10,570					29	78	82	17	106	95	22	26	116	127	125	97	79	120	127
120	Clairlea-Birchmount	18,795					30	28	25	97	78	51	24	71	64	22	25	78	102	25	20
47	Don Valley Village	18,285					31	27	27	9	19	38	42	43	101	53	65	52	77	23	32
14	Islington-City Centre West	30,735					32	6	5	52	26	5	106	47	40	5	7	8	69	2	8
44	Flemingdon Park	14,580					33	46	49	12	61	78	28	103	128	57	100	72	140	51	51
106	Humewood-Cedarvale	10,265					34	90	88	106	74	71	111	62	11	120	119	139	21	138	122
26	Downsview-Roding-CFB	23,765					35	9	12	29	66	20	47	132	80	11	16	43	99	5	5
133	Centennial Scarborough	8,835					36	98	99	104	131	97	12	19	68	131	124	133	71	135	133
108	Briar Hill-Belgravia	10,320					37	92	86	30	74	108	73	107	50	80	47	86	106	108	82
25	Glenfield-Jane Heights	19,715					38	17	22	18	91	40	13	138	136	39	37	93	126	16	13
16	Stonegate-Queensway	16,815					39	32	36	96	78	27	88	34	43	49	34	56	27	58	58
2	Mount Olive-Silverstone-Jamestown	22,330					40	13	16	21	113	45	4	134	129	85	84	101	137	37	15

Figure 4: Results dataset can be downloaded from Supplemental Materials [1] at the end of the report.

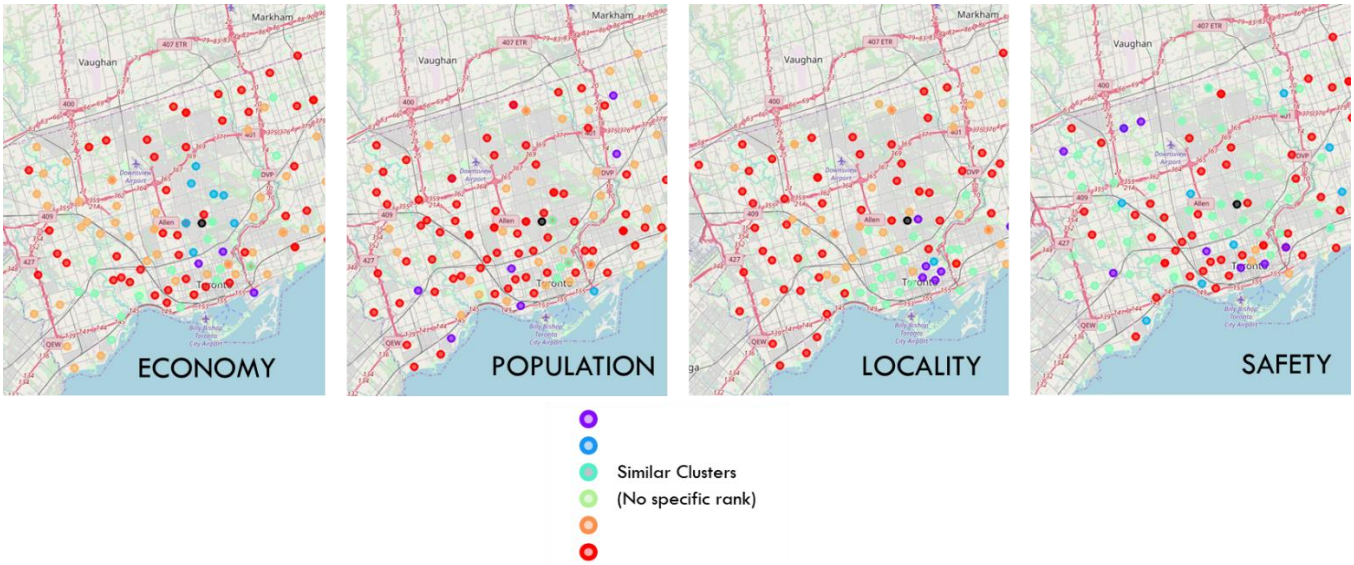

Figure 5: Clusters of similar neighborhoods based on different variable groups.

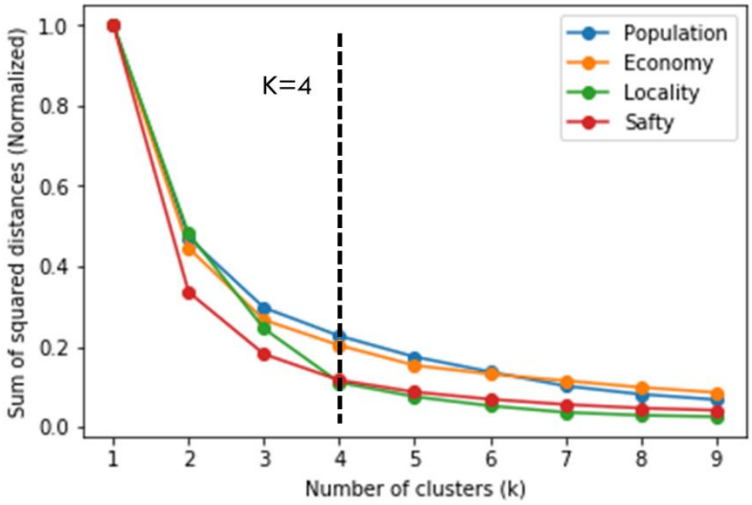

Figure 6: k=4 was determined to be the optimum k value using "Sum of squared distances"

To calculate the overall rank, the values of *a =1* and *b =2*. Client can run the overall rank based on their weights (*a* and *b*) of positive and negative variables based on their expectations.

Figure 7: (a) Neighborhood ranks. (b) Exploring top ranked neighborhood.

Discussion

By analyzing demographics of inhabitants, economy, locality, and other factors, there were many interesting patters and trends were emerged out. Those can be useful making business decisions and form strategies in opening/running a business. This study can be further extended to other applications or create a generalized by analyzing the neighborhoods with different perspectives using available data. Clustering is used in the present work. We can also utilize other machine learning methods and algorithms to build robust prescriptive models. The current analysis can be farther strength by having few iterations with the client going through the analysis results and changing the perspective of analysis per client expectations.

Conclusion

This work is mainly focused on creating a usable dataset and exploratory analysis which will help the client to know personalities of the neighborhoods they are considering. Similar neighborhoods were grouped using an unsupervised machine learning method K-means clustering. The client is given functions to run the analysis and clustering based on their needs as well as explore selected neighborhoods. Client can use the outcome of this project to have better understanding on similar neighborhoods that they can form limited number of business strategies and models.

Supplemental Materials

1. Results dataset, https://github.com/sptennak/Coursera_Capstone/blob/master/NeighbourhoodDataSet_Ranked.xlsx

References

1. City of Toronto's Open Data Catalogue, https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/.
2. IBM Data Science Professional Certificate Course materials and assignments, https://www.coursera.org/specializations/ibm-data-science-professional-certificate.
3. Foursquare API documentation, https://developer.foursquare.com
4. Python Data Analysis Library, https://pandas.pydata.org/
5. Wikipedia pages as a data source, https://en.wikipedia.org/wiki/List_of_city-designated_neighbourhoods_in_Toronto, https://en.wikipedia.org/wiki/ [search key].

Appendix

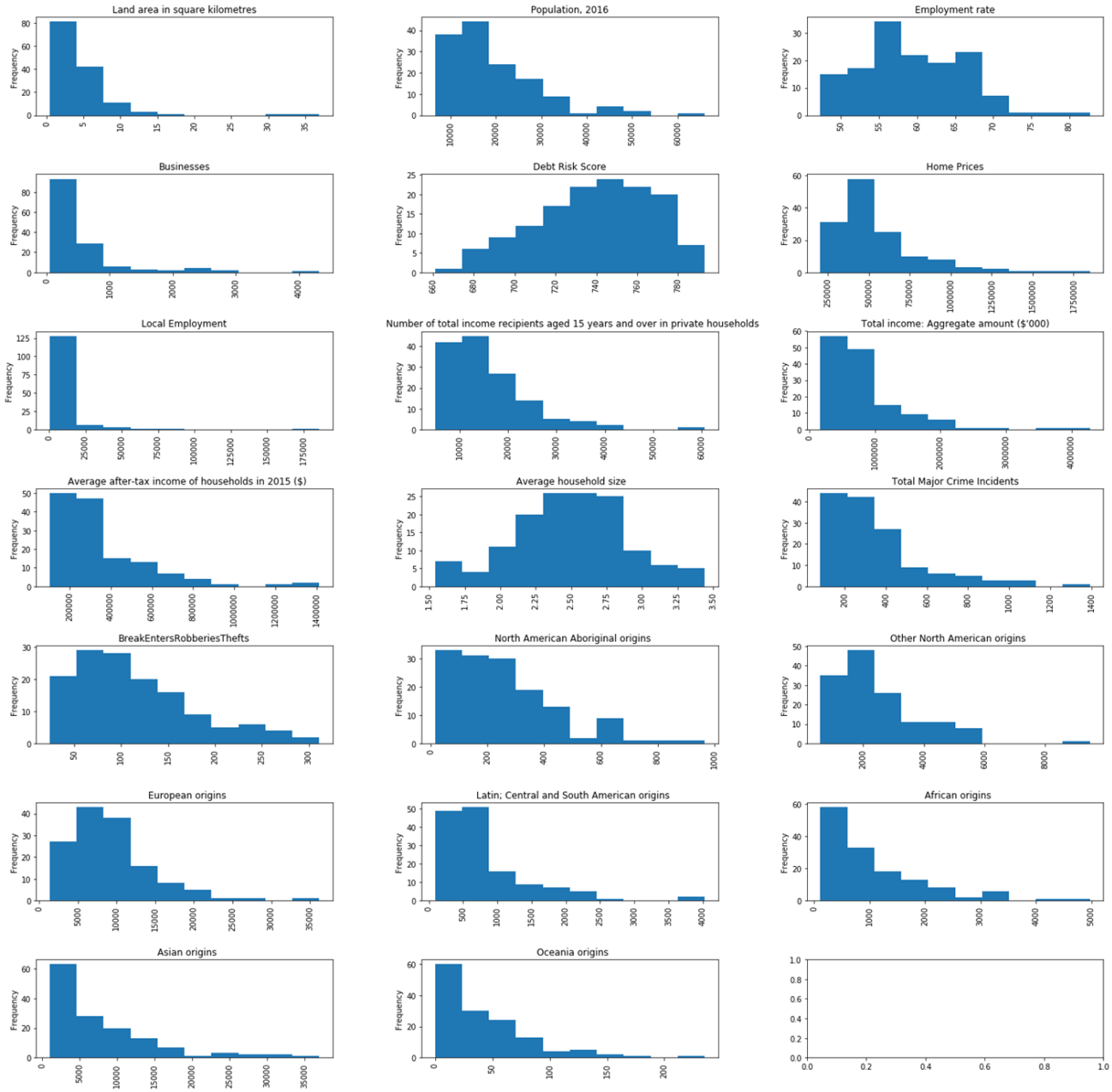

Figure A1: Distributions of neighborhood variables.

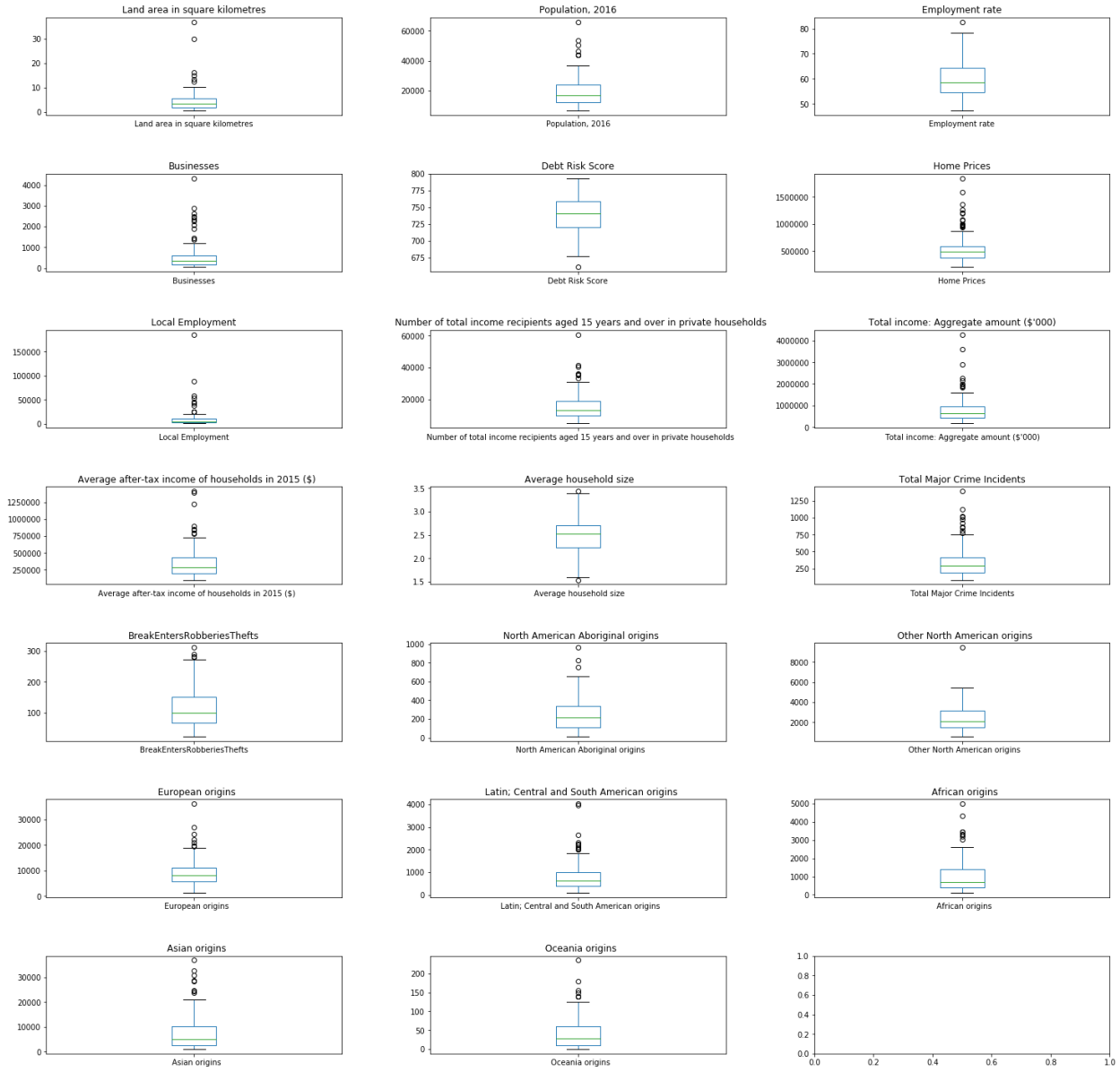

Figure A2: Distributions of neighborhood variables using box plot.

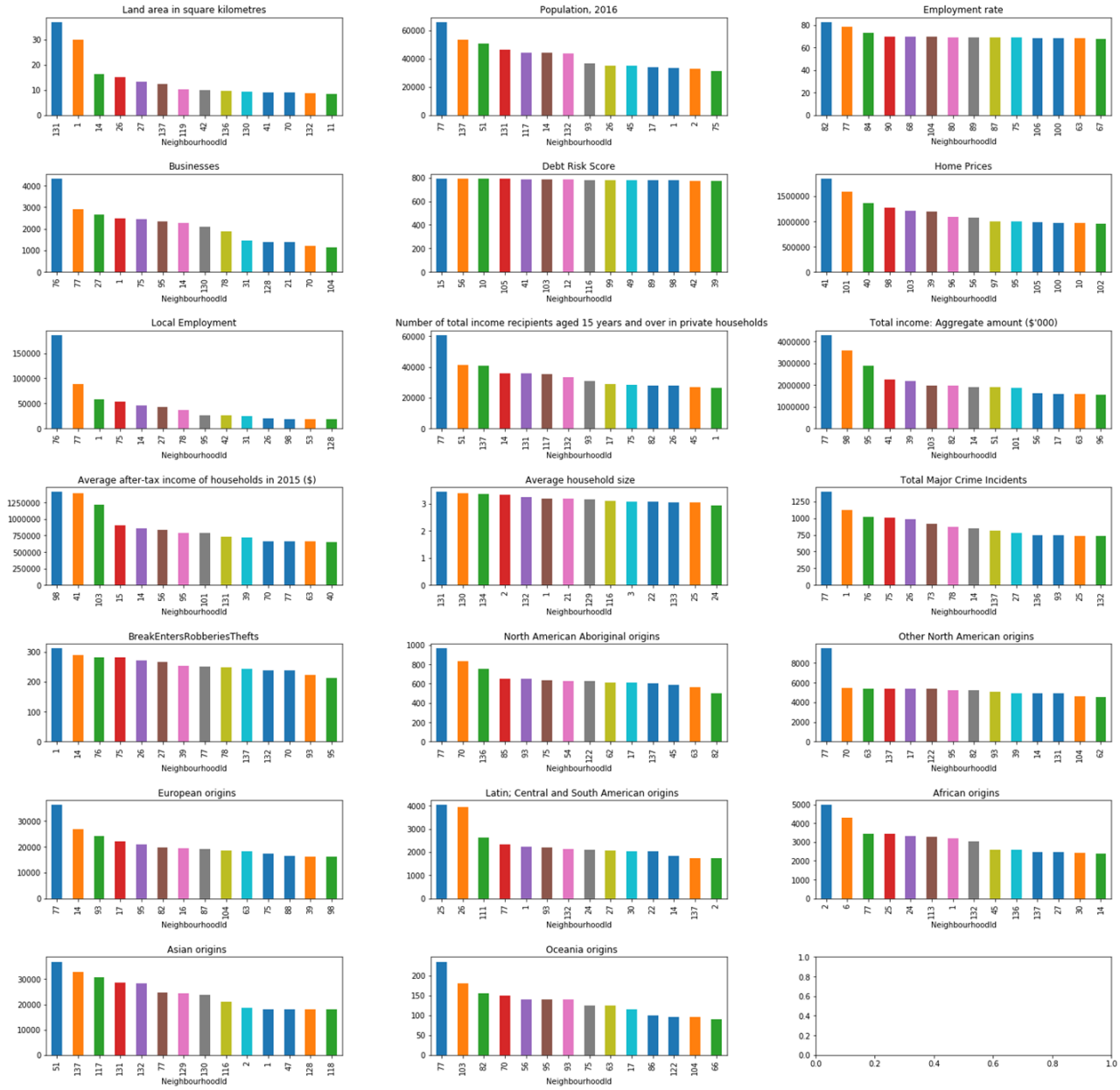

Figure A3: Top 10% (14/140) neighborhoods for each variable.

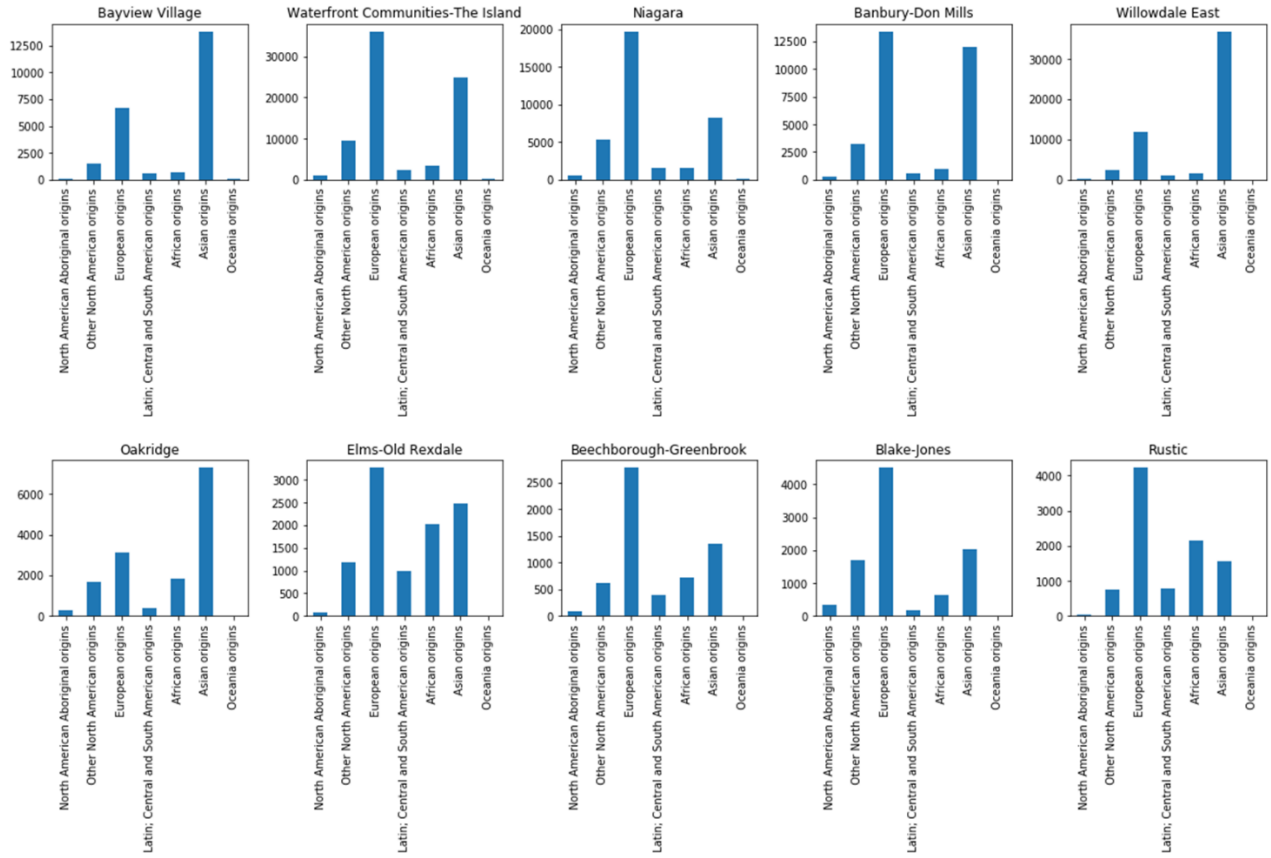

Figure A4: Population origins of selected neighborhoods.

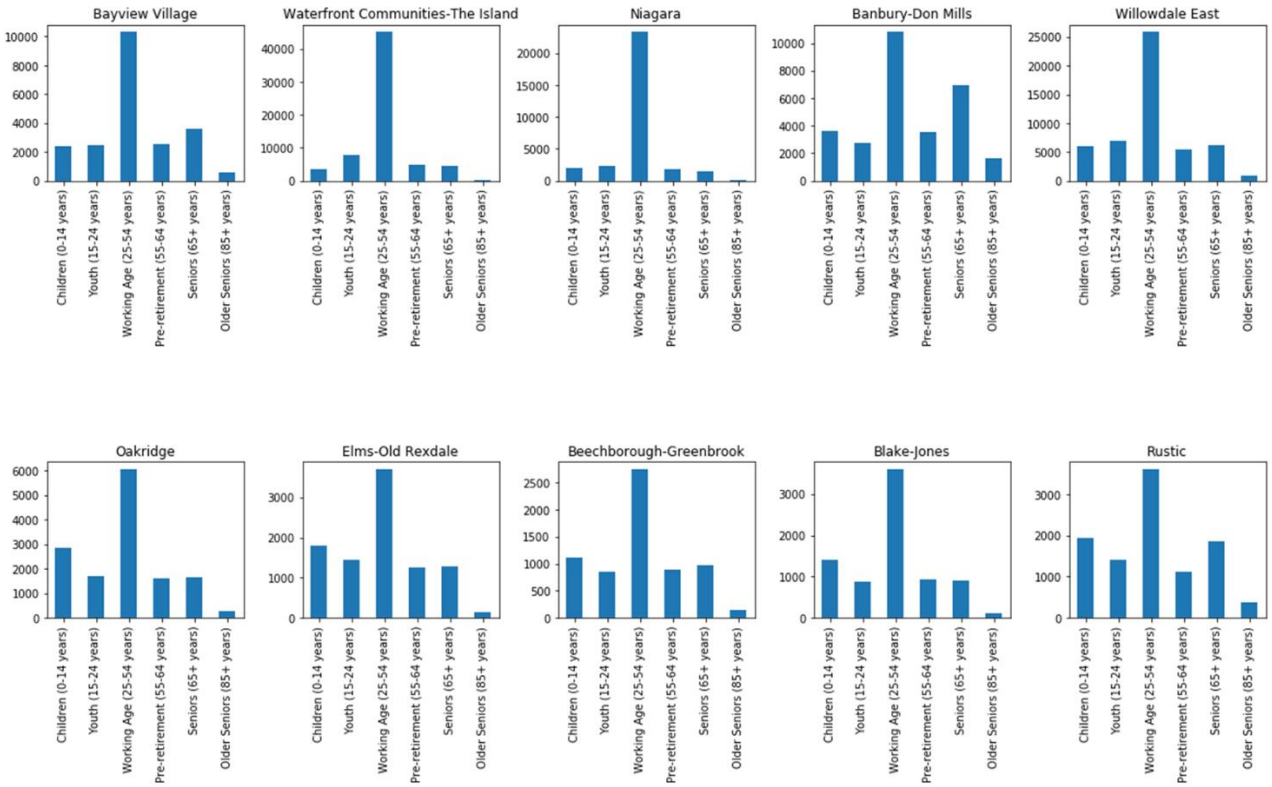

Figure A5: Age group distribution of selected neighborhoods.

Table: Clustering results

			Cluster			
Neighborhood Id	**Neighborhood**	**Target Population**	**Population**	**Economy**	**Locality**	**Safety**
1	West Humber-Clairville	23,280	4	0	0	4
2	Mount Olive-Silverstone-Jamestown	22,330	4	0	4	1
3	Thistletown-Beaumond Heights	6,765	0	4	0	3
4	Rexdale-Kipling	7,175	0	4	0	3
5	Elms-Old Rexdale	6,395	0	4	0	3
6	Kingsview Village-The Westway	14,205	4	0	0	0
7	Willowridge-Martingrove-Richview	13,670	4	0	0	0
8	Humber Heights-Westmount	6,440	0	4	0	3
9	Edenbridge-Humber Valley	10,130	0	3	0	3
10	Princess-Rosethorn	7,260	0	0	0	3
11	Eringate-Centennial-West Deane	12,170	0	0	0	3
12	Markland Wood	6,500	0	4	0	3
13	Etobicoke West Mall	8,085	0	4	4	3
14	Islington-City Centre West	30,735	1	3	0	1
15	Kingsway South	5,690	0	3	0	3
16	Stonegate-Queensway	16,815	4	3	0	0
17	Mimico (includes Humber Bay Shores)	25,325	1	3	0	2
18	New Toronto	8,185	0	4	0	0
19	Long Branch	7,355	0	4	0	3
20	Alderwood	8,280	0	4	0	3

21	Humber Summit	7,890	0	4	0	0
22	Humbermede	10,430	0	4	0	0
23	Pelmo Park-Humberlea	7,405	0	4	0	3
24	Black Creek	14,240	4	4	4	2
25	Glenfield-Jane Heights	19,715	4	0	4	1
26	Downsview-Roding-CFB	23,765	1	0	0	4
27	York University Heights	20,005	4	0	0	1
28	Rustic	6,135	0	4	0	3
29	Maple Leaf	6,580	0	4	0	3
30	Brookhaven-Amesbury	12,270	0	4	4	0
31	Yorkdale-Glen Park	9,540	0	4	0	2
32	Englemount-Lawrence	13,845	4	0	4	0
33	Clanton Park	11,315	0	0	0	0
34	Bathurst Manor	10,625	0	0	0	3
35	Westminster-Branson	17,720	4	0	4	0
36	Newtonbrook West	16,845	4	0	4	0
37	Willowdale West	11,780	0	0	4	3
38	Lansing-Westgate	11,605	0	0	0	3
39	Bedford Park-Nortown	14,695	4	2	0	0
40	St.Andrew-Windfields	11,965	0	2	0	0
41	Bridle Path-Sunnybrook-York Mills	6,040	0	2	0	3
42	Banbury-Don Mills	17,095	4	3	0	0
43	Victoria Village	11,470	0	4	0	3

44	Flemingdon Park	14,580	4	4	3	0
45	Parkwoods-Donalda	23,725	1	3	0	0
46	Pleasant View	10,570	0	4	4	3
47	Don Valley Village	18,285	4	0	4	2
48	Hillcrest Village	10,495	0	0	0	3
49	Bayview Woods-Steeles	7,950	0	0	0	3
50	Newtonbrook East	11,200	0	0	0	0
51	Willowdale East	38,250	1	3	3	2
52	Bayview Village	15,355	4	0	0	3
53	Henry Farm	11,765	0	4	4	3
54	O'Connor-Parkview	12,890	0	0	0	0
55	Thorncliffe Park	13,140	4	4	4	3
56	Leaside-Bennington	10,970	0	2	0	3
57	Broadview North	8,155	0	4	4	3
58	Old East York	6,230	0	4	0	3
59	Danforth East York	11,515	0	0	4	3
60	Woodbine-Lumsden	5,615	0	4	4	3
61	Taylor-Massey	11,020	0	4	1	0
62	East End-Danforth	15,005	4	0	4	2
63	The Beaches	14,805	4	3	4	0
64	Woodbine Corridor	8,825	0	0	4	3
65	Greenwood-Coxwell	10,310	0	4	3	0
66	Danforth	6,485	0	4	3	0

67	Playter Estates-Danforth	5,430	0	0	3	3
68	North Riverdale	8,305	0	0	4	0
69	Blake-Jones	5,430	0	4	4	3
70	South Riverdale	20,590	4	3	0	1
71	Cabbagetown-South St. James Town	8,605	0	0	3	0
72	Regent Park	8,450	0	4	1	3
73	Moss Park	17,085	4	0	1	1
74	North St. James Town	14,585	3	4	2	0
75	Church-Yonge Corridor	27,075	3	3	1	4
76	Bay Street Corridor	21,685	3	3	1	4
77	Waterfront Communities-The Island	57,625	2	1	3	4
78	Kensington-Chinatown	13,905	4	0	3	1
79	University	5,725	0	4	4	0
80	Palmerston-Little Italy	10,570	0	0	3	0
81	Trinity-Bellwoods	12,465	0	0	3	0
82	Niagara	27,620	1	3	3	0
83	Dufferin Grove	8,980	0	4	3	0
84	Little Portugal	12,315	0	0	3	0
85	South Parkdale	17,190	4	0	3	2
86	Roncesvalles	10,985	0	0	3	0
87	High Park-Swansea	16,565	4	3	0	3
88	High Park North	16,135	4	3	3	0
89	Runnymede-Bloor West Village	6,780	0	0	4	3

90	Junction Area	10,685	0	0	4	0
91	Weston-Pelham Park	8,105	0	4	4	0
92	Corso Italia-Davenport	10,215	0	4	4	0
93	Dovercourt-Wallace Emerson-Junction	27,650	1	3	3	1
94	Wychwood	9,335	0	0	3	3
95	Annex	22,270	4	1	3	2
96	Casa Loma	7,100	0	2	4	3
97	Yonge-St.Clair	8,420	0	3	3	3
98	Rosedale-Moore Park	13,140	4	1	0	3
99	Mount Pleasant East	11,345	0	3	4	3
100	Yonge-Eglinton	8,410	0	0	4	3
101	Forest Hill South	7,120	0	2	0	3
102	Forest Hill North	8,615	0	0	4	3
103	Lawrence Park South	10,080	0	2	0	3
104	Mount Pleasant West	22,650	3	3	1	0
105	Lawrence Park North	9,375	0	3	4	3
106	Humewood-Cedarvale	10,265	0	0	4	3
107	Oakwood Village	14,400	4	0	3	0
108	Briar Hill-Belgravia	10,320	0	4	4	3
109	Caledonia-Fairbank	7,140	0	4	4	3
110	Keelesdale-Eglinton West	7,655	0	4	4	0
111	Rockcliffe-Smythe	15,085	4	0	0	0
112	Beechborough-Greenbrook	4,490	0	4	0	3

113	Weston	12,440	0	4	4	2
114	Lambton Baby Point	5,265	0	0	0	3
115	Mount Dennis	9,505	0	4	4	0
116	Steeles	15,700	4	0	4	3
117	L'Amoreaux	28,870	1	3	4	0
118	Tam O'Shanter-Sullivan	17,410	4	0	4	0
119	Wexford/Maryvale	18,970	4	0	0	2
120	Clairlea-Birchmount	18,795	4	0	0	2
121	Oakridge	9,325	0	4	4	0
122	Birchcliffe-Cliffside	15,035	4	0	0	0
123	Cliffcrest	10,660	0	0	0	0
124	Kennedy Park	11,805	0	4	0	2
125	Ionview	9,415	0	4	4	3
126	Dorset Park	16,970	4	4	0	0
127	Bendale	20,190	4	0	0	2
128	Agincourt South-Malvern West	16,590	4	0	0	0
129	Agincourt North	19,240	4	0	0	0
130	Milliken	17,790	4	0	0	0
131	Rouge	31,900	1	3	0	2
132	Malvern	30,020	1	0	0	1
133	Centennial Scarborough	8,835	0	0	0	3
134	Highland Creek	8,565	0	4	0	3
135	Morningside	11,885	0	4	0	0

136	West Hill	18,500	4	0	0	1
137	Woburn	35,850	1	3	0	1
138	Eglinton East	15,135	4	4	4	2
139	Scarborough Village	11,140	0	4	4	2
140	Guildwood	5,990	0	4	0	3