FINDING OPTIMAL LOCATION TO OPEN RESTAURANT/GROCERY BUSINESS

APPLIED DATA SCIENCE CAPSTONE

THE BATTLE OF NEIGHBORHOODS

Sumudu Tennakoon

2018-01-02

CONTENT

- Introduction
- Problem understanding
- Methodology
- Source data
- Exploratory data analysis
- Neighborhood clustering
- Neighborhood Ranks
- Discussion
- Conclusion

INTRODUCTION

- 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.

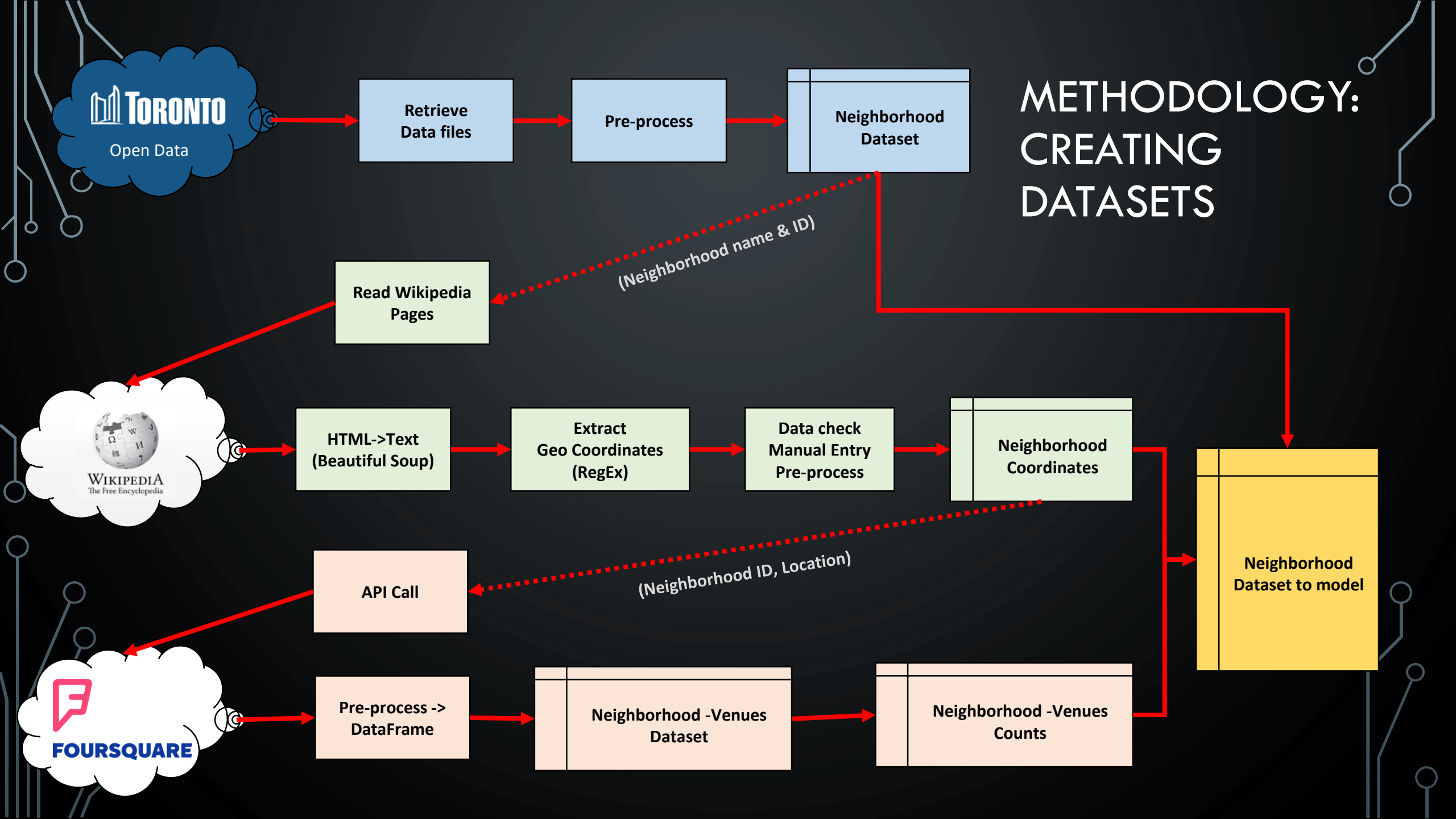
PROBLEM UNDERSTANDING

- 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 140 postal zip codes assigned to 103 different boroughs.
- 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.

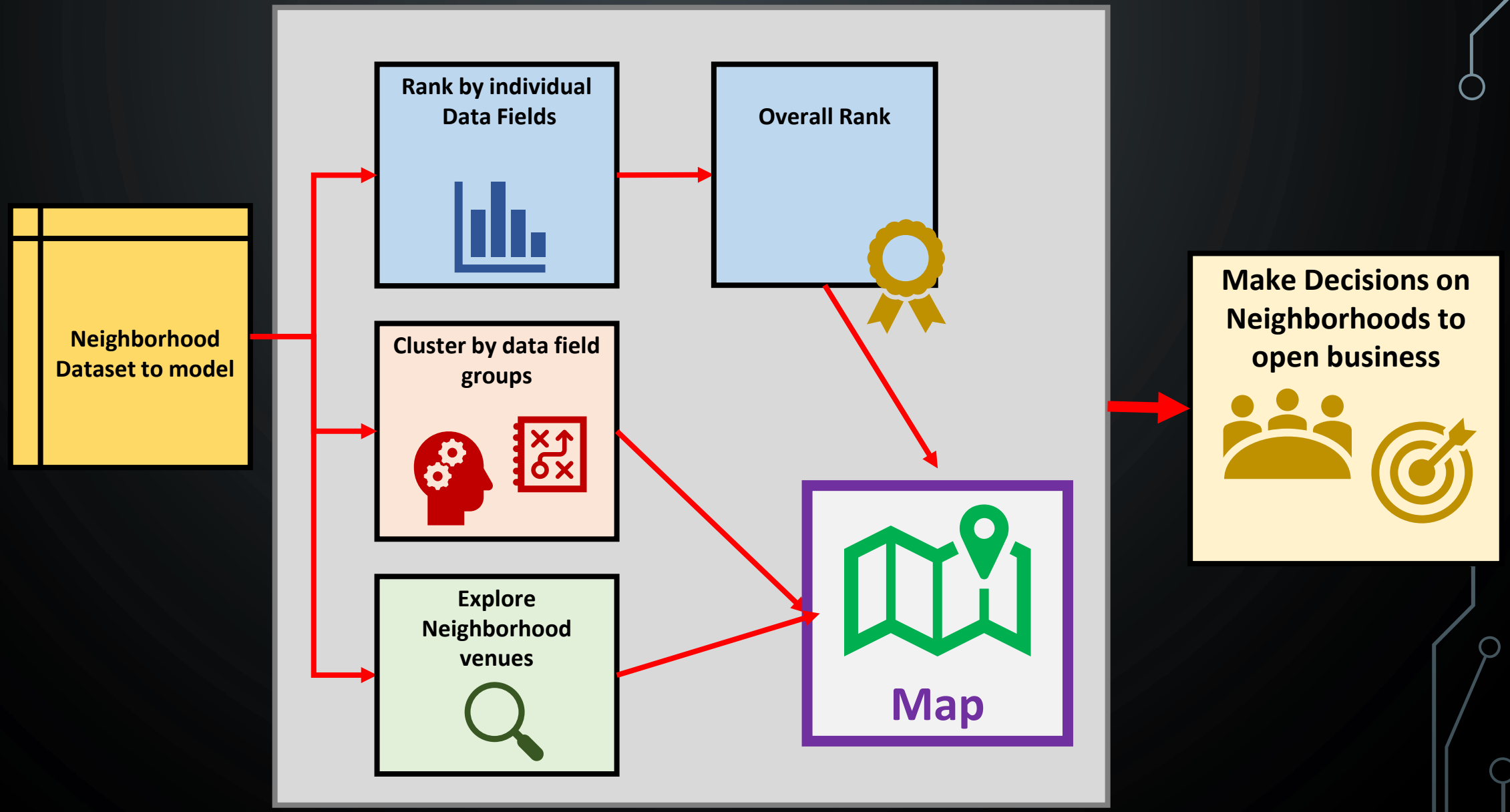
METHODOLOGY

- 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.

METHODOLOGY: CREATING DATASETS



METHODOLOGY



SOURCE DATA

Source #1: City of Toronto's Open Data Catalogue

- URL: https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/
- 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.

SOURCE DATA

Source #2: Geospatial Coordinates

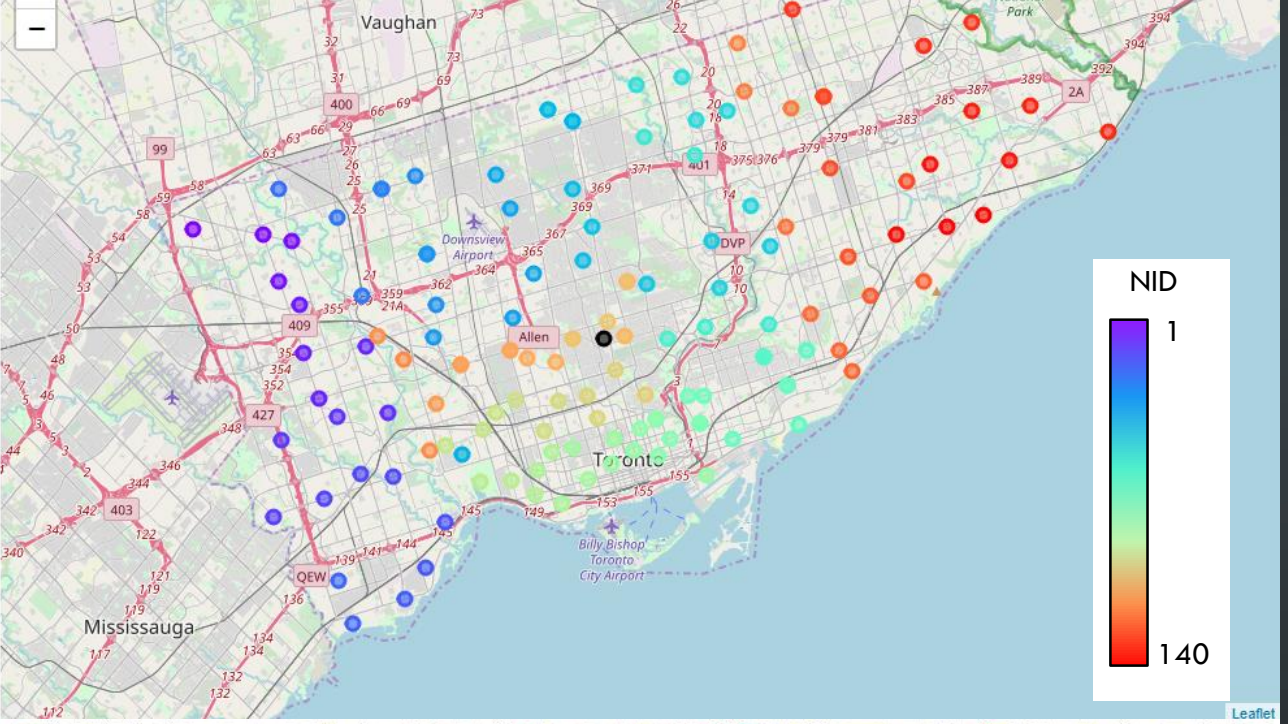
URL: https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/

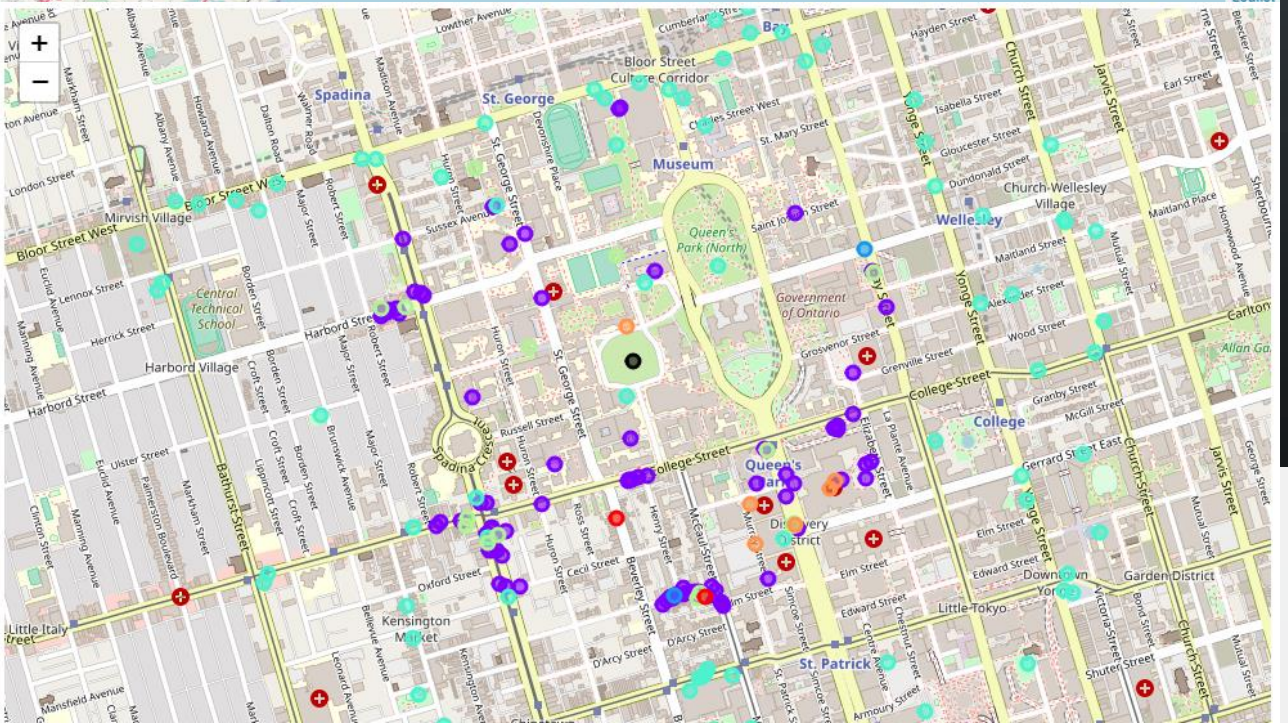
- Web scraping were done across number of Wikipedia pages starting from https://en.wikipedia.org/wiki/List_of_city-designated_neighbourhoods_in_Toronto. The geo locations were extracted from each Neighborhood's Wikipedia page and from few of them had to do manual data entry after referring to other sources.

Source #3: Foursquare APIs location data

- URL: https://developer.foursquare.com
- The foursquare dataset will be used to identify competitive business locations in each neighborhood (e.g. Grocery stores, restaurants) as well as venues which adds new businesses opportunities (e.g. Schools, Offices, Attractions, Shopping Malls, etc.).

NEIGHBOURHOODS

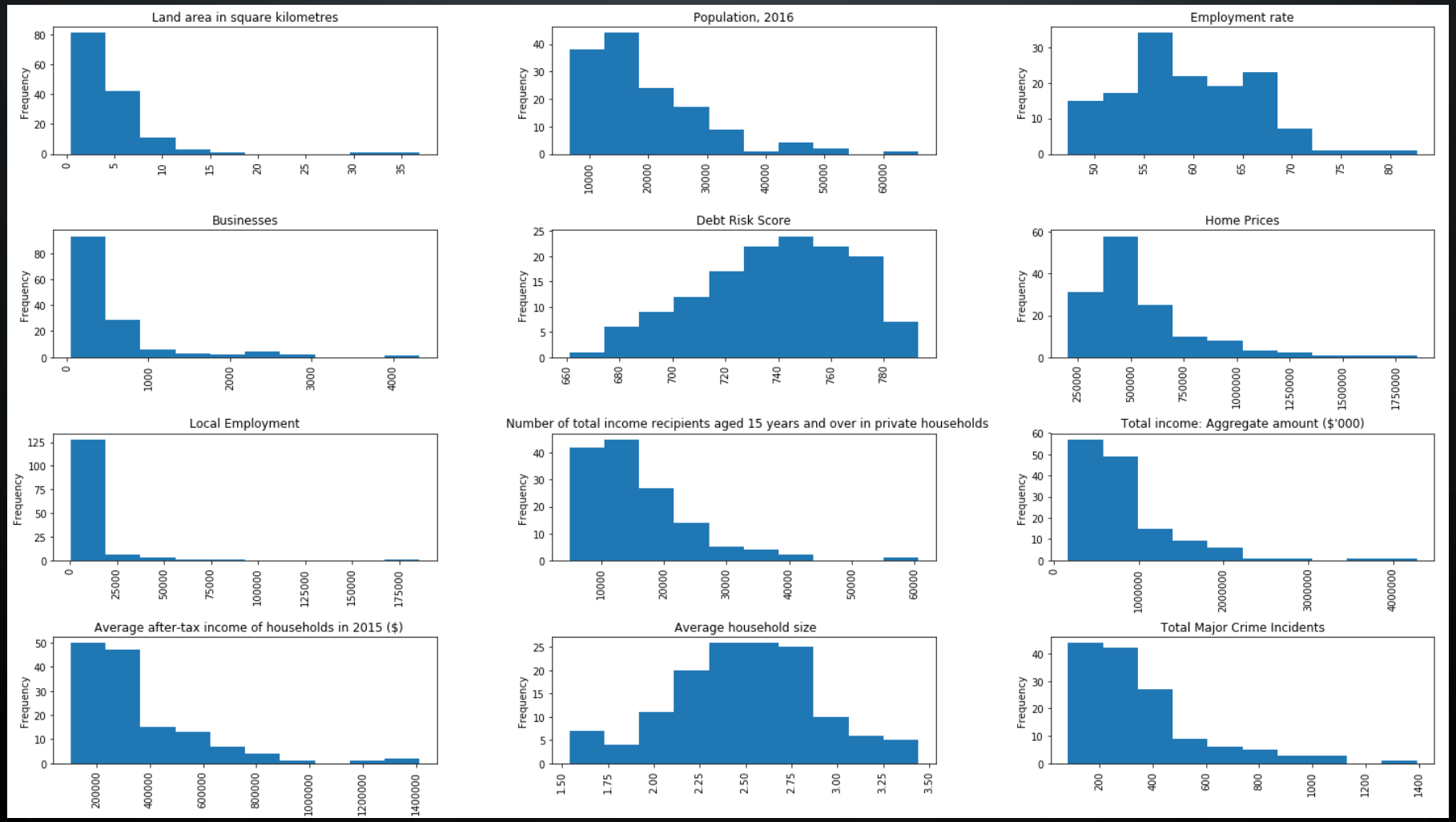




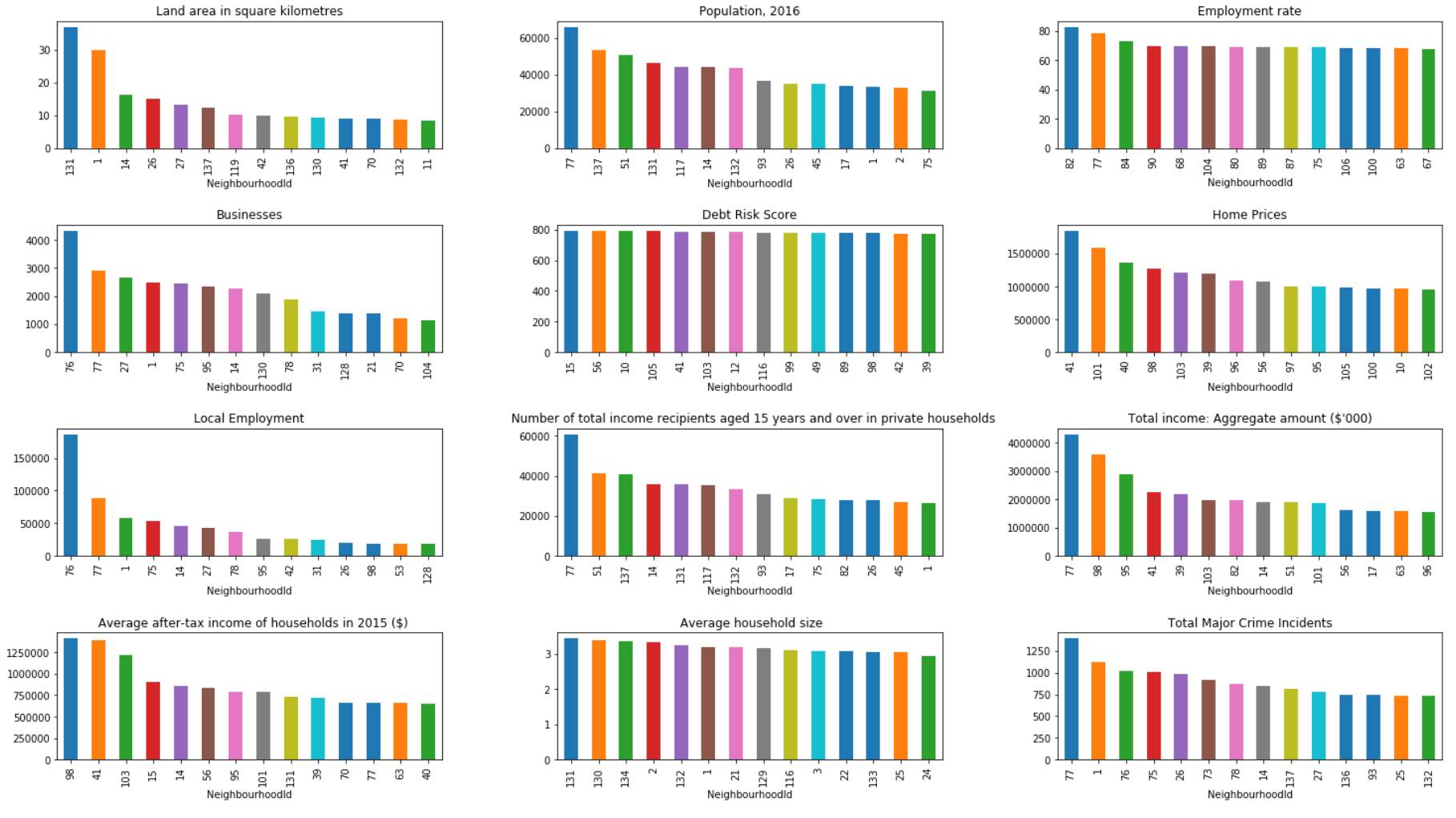
SOFTWARE PLATFORM & TOOLS

- 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/)

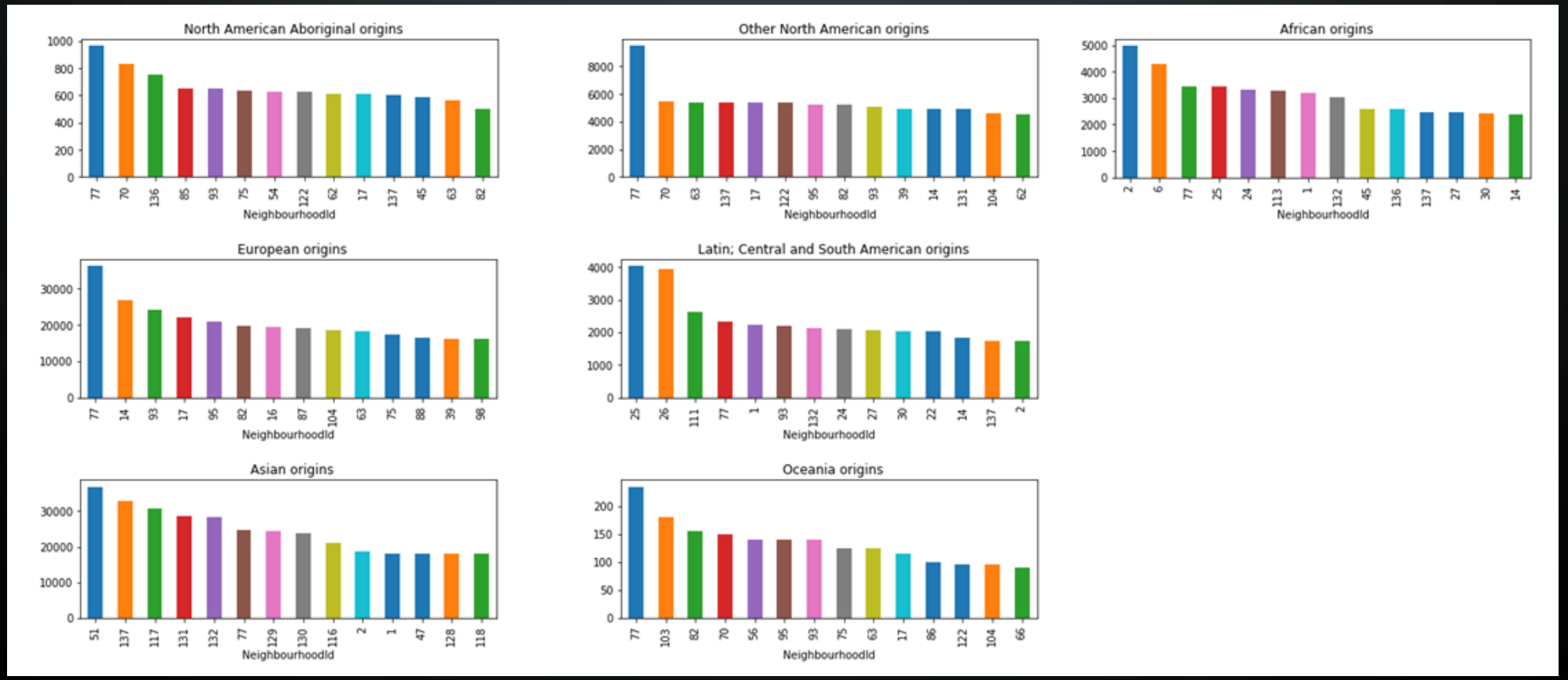
EXPLORATORY DATA ANALYSIS



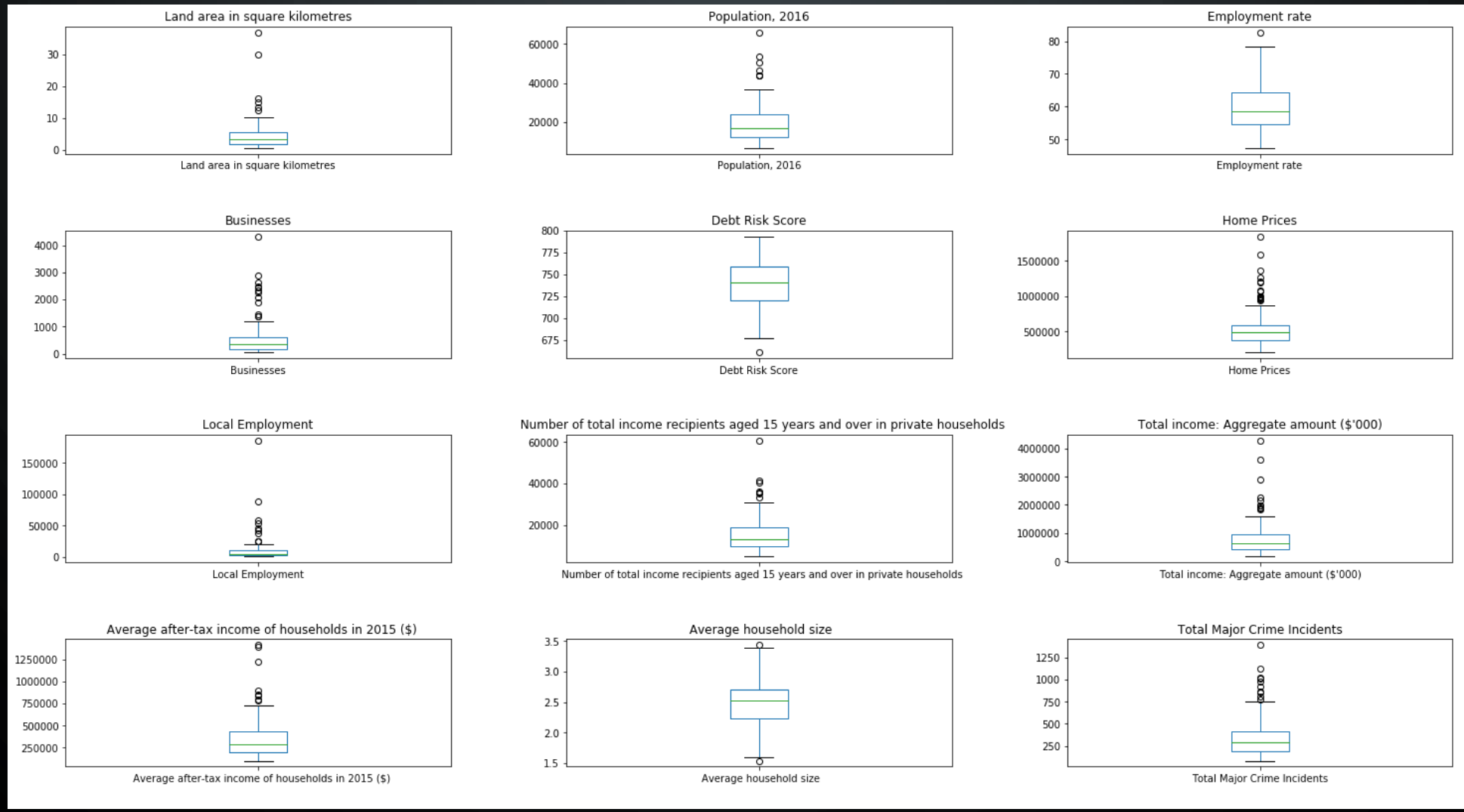
EXPLORATORY DATA ANALYSIS



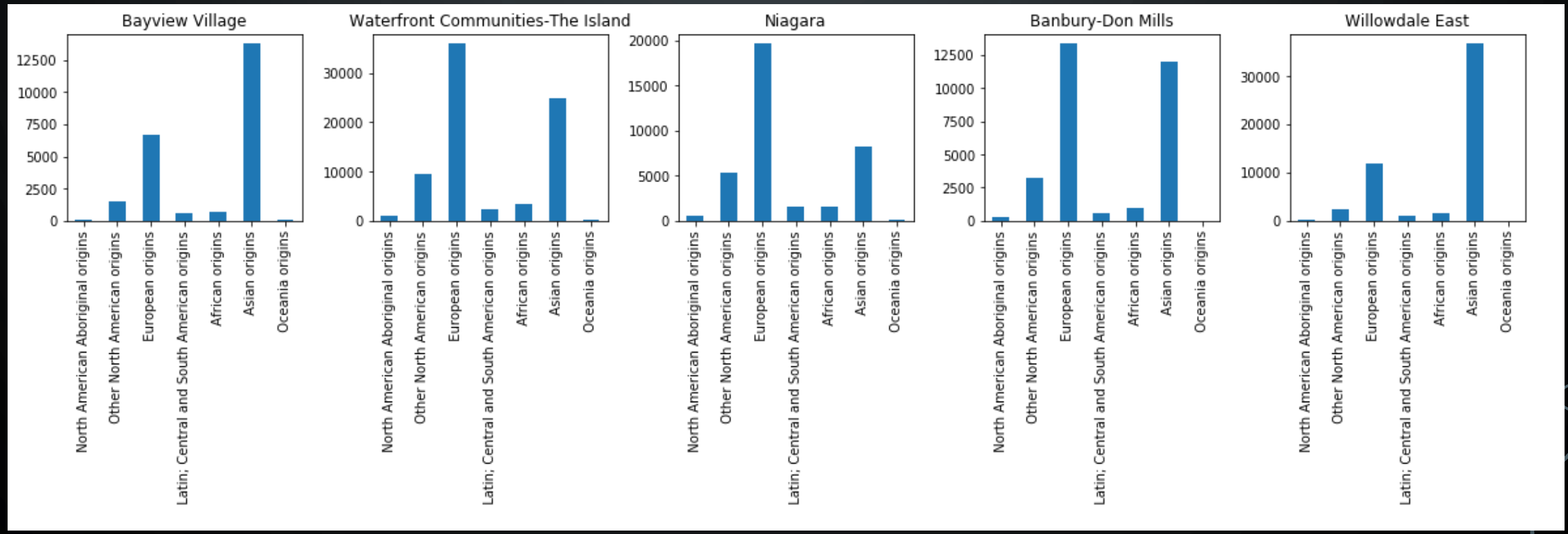
EXPLORATORY DATA ANALYSIS (ORIGINS)



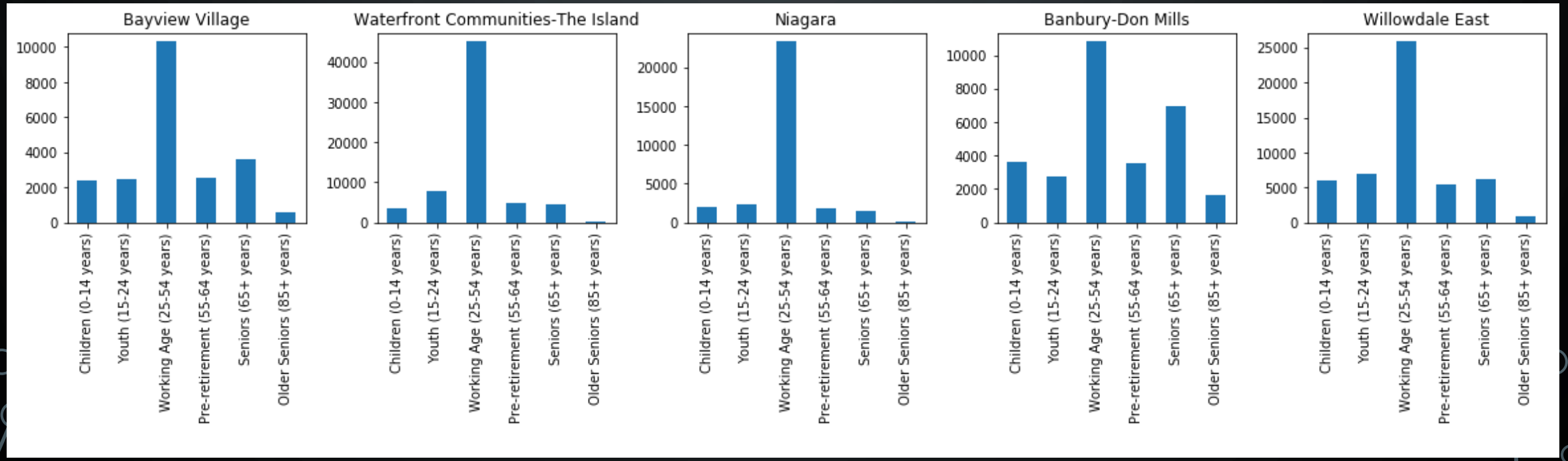
EXPLORATORY DATA ANALYSIS

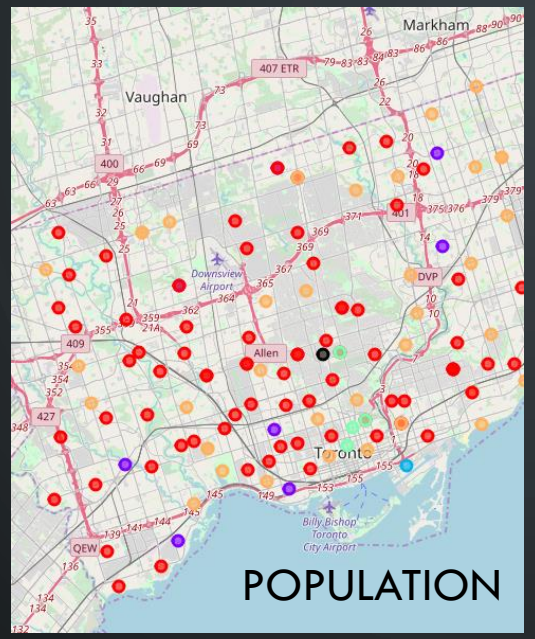


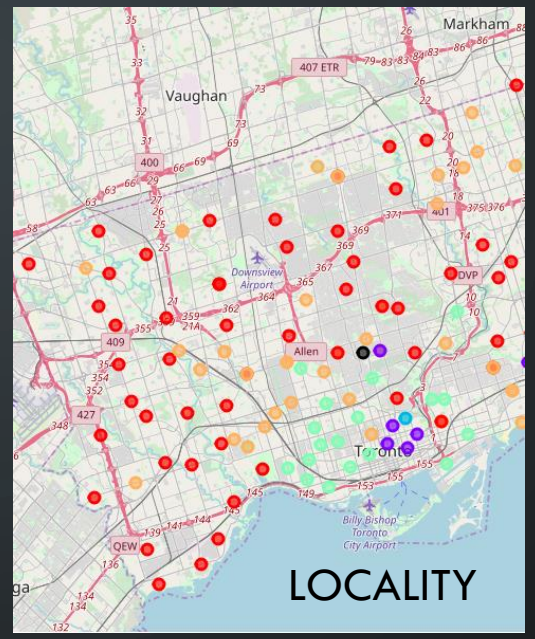
EXPLORATORY DATA ANALYSIS

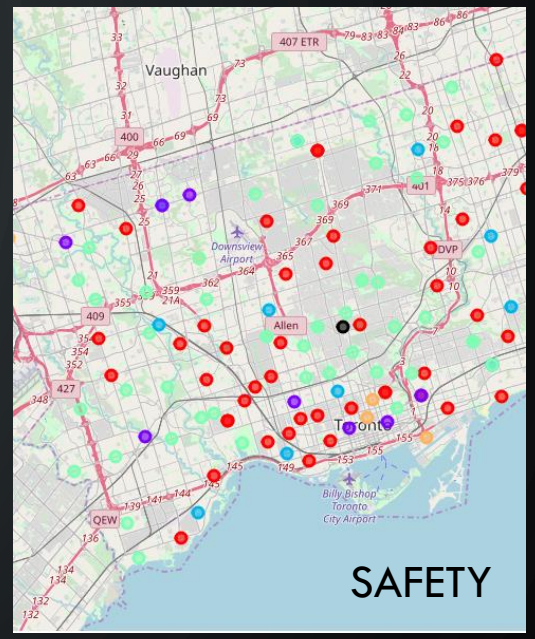


EXPLORATORY DATA ANALYSIS





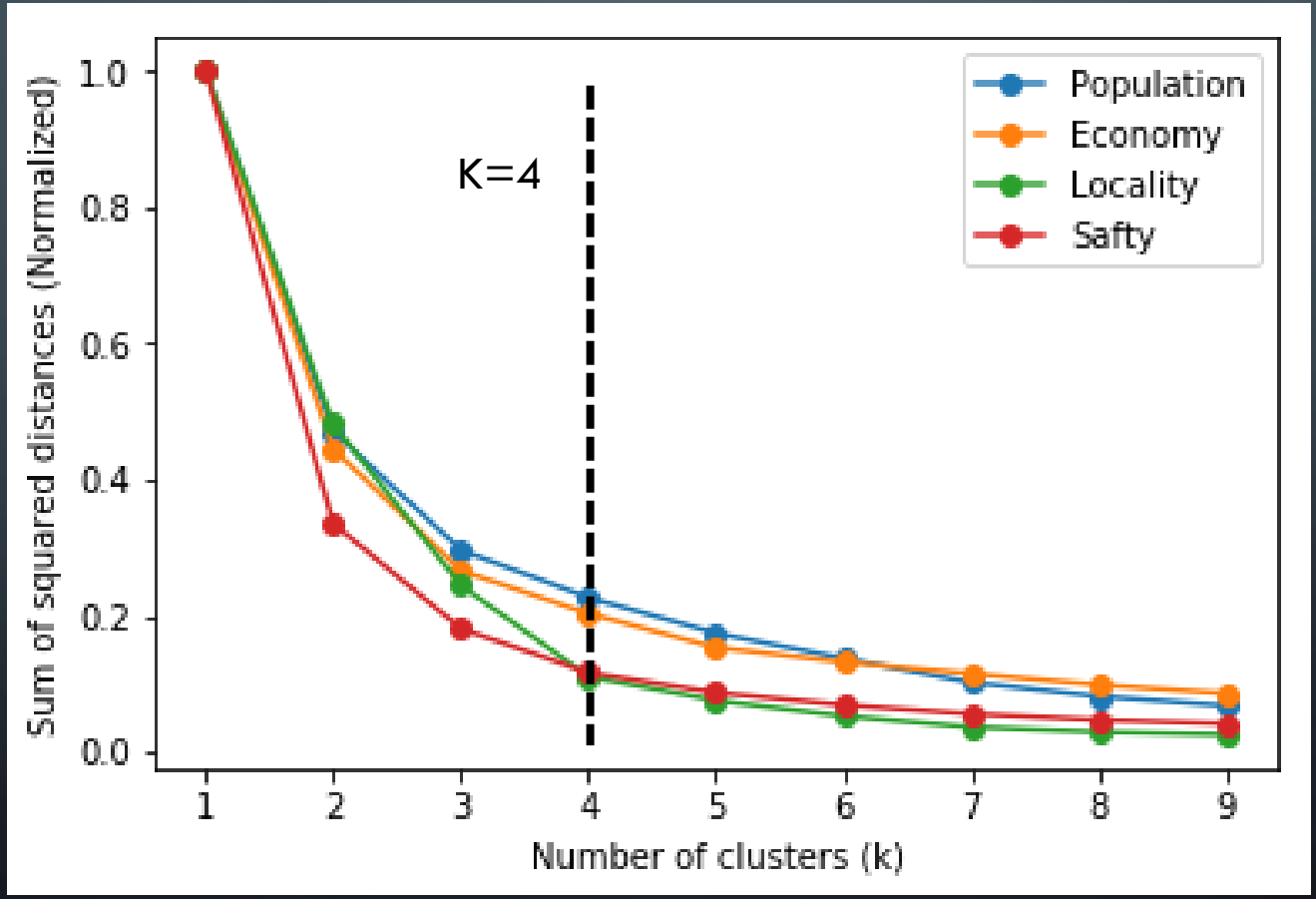




NEIGHBORHOOD CLUSTERING

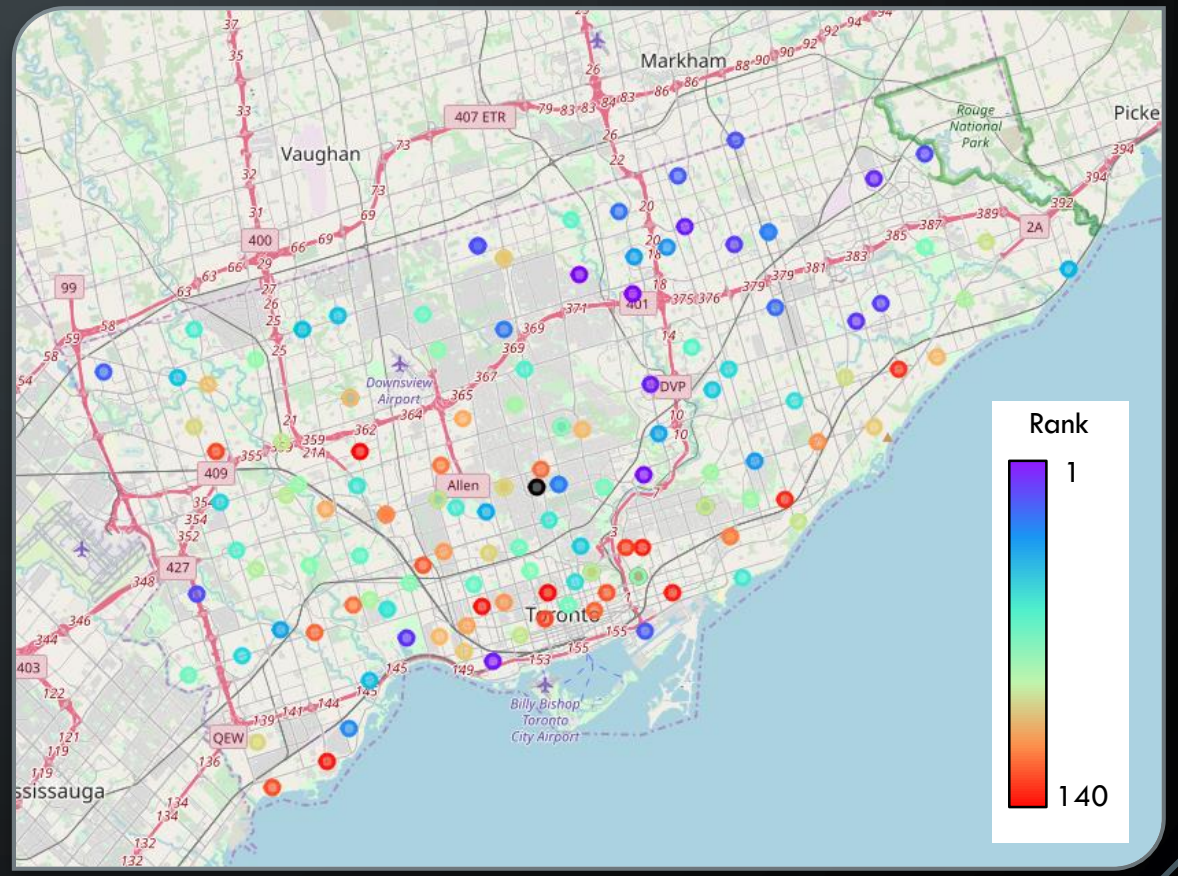
BY DIFFERENT FACTORS

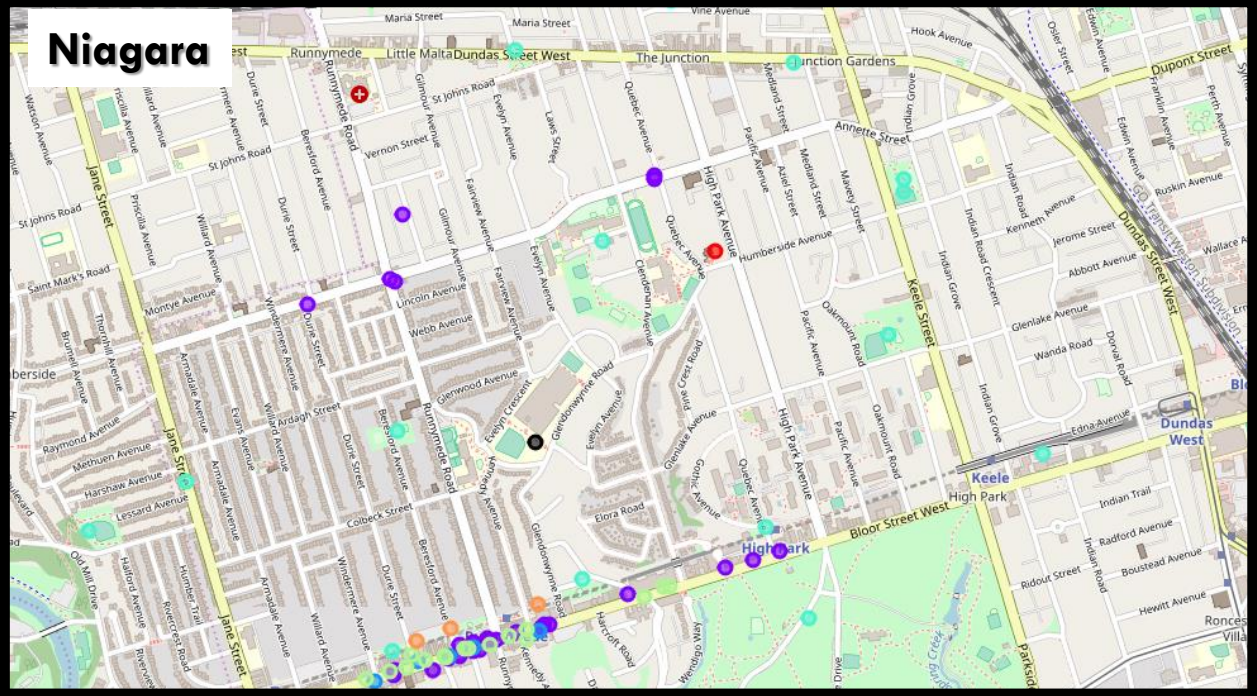
FINDING BEST K USING "SUM OF SQUARED DISTANCES"

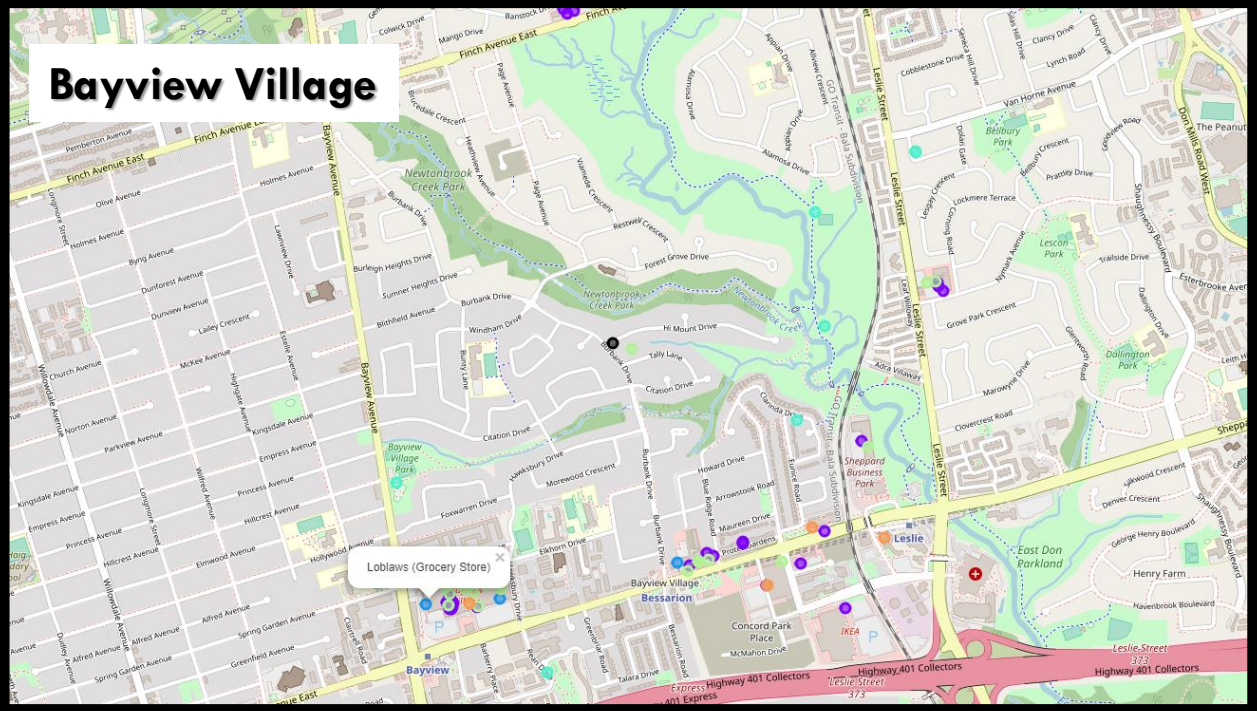


NEIGHBORHOOD RANKS

- Neighborhoods were ranked based on business potentials identified by the project.







EXPLORE HIGH RANKED NEIGHBORHOODS

RESULTS DATASET

				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	CompetitiveVenues	Home Prices	BreakEntersRobberiesTheffs	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

EXPLORE HIGH RANKED NEIGHBORHOODS

- Niagara

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

CONCLUSION

- This work is mainly focused on creating a usable dataset and explanatory 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.
- Client is given functions to run the analysis and clustering based on their needs as well as explore selected neighborhoods.
- The 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.

REFERENCES

- City of Toronto's Open Data Catalogue, https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/.
- IBM Data Science Professional Certificate Course materials and assignments, https://www.coursera.org/specializations/ibm-data-science-professional-certificate.
- Foursquare API documentation, https://developer.foursquare.com
- Python Data Analysis Library, https://pandas.pydata.org/
- 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].