Applied Data Science Capstone Project – The Battle of Neighbourhoods

Baltej Toor

INTRODUCTION

Background

Given its growing population and corresponding increased development and investment into the city, Toronto provides an exceptional opportunity to grow a new business. Regardless of the numerous opportunities in the GTA, there are important factors associated with starting any new venture such as determining a location suited to success. When deciding on a location, essential aspects of the area include security/safety, marketing and promotion opportunities, and overall business concerns such as organized capital improvements and planning. The aforementioned considerations need to be made in order to set a small business up to thrive in a new environment. However, there are resources available to small business owners to establish a setting in which businesses can operate in a safe and competitive climate. A primary example of such a resource is Toronto's BIA program. A BIA (Business Improvement Area) is an association of commercial property owners and tenants that work together in a partnership, defined in a given area, with the city of Toronto to create an atmosphere in which business can be conducted safely and competitively in such a way that attracts customers and new businesses. BIAs in many ways are the catalysts for the improvements that create the kinds of strong communities that allow new businesses to integrate into and contribute to Toronto's economy.

Business Problem

Despite the opportunities and resources afforded to new businesses, there are difficulties that face entrepreneurs and business owners looking to grow their operation in one of Toronto's neighbourhoods. Depending on the nature of the business it may be challenging to secure enough capital towards efforts to gather extensive information on competitors and BIA initiatives. Entrepreneurs typically enter a market with limited resources, the majority of which needing to go towards inventory, leases/rental costs, and labour. This project aims to provide a tool in which new business owners can evaluate the level of competition in each Toronto neighbourhood and cross-reference these locations against established BIAs. Businesses will be able to evaluate locations on the basis of competition density and enhancement via the Business Improvement Areas.

DATA

In order to determine the viability of a particular area a combination of Foursquare location data and the Toronto BIA open dataset will be utilized. The two sources will be used in conjunction to determine the density of competition on the neighbourhood level as well as in regards to a common specified radius.

Foursquare Data

Data Setup

In order to take advantage of the Foursquare API's venue exploration feature we'll need neighbourhood geolocation data. Geospatial coordinates data for Toronto neighbourhoods will be read from an external dataset (via http://cocl.us/Geospatial_data - geospatial_coordinates.csv). This data will be cross-referenced with postal code, borough and neighbourhood data from a listing of neighbourhoods (via resources that facilitate web scraping through the link https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). After modifications to concatenate neighbourhood values and ignore incomplete records, the foundational dataset with which the Foursquare API call is made will look similar to the following (Figure 1):

	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	Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494
...	...	...	...	...	...

Figure 1 *Constructed dataframe to be used for Foursquare API call including fields PostalCode, Borough, Neighbourhood, Latitude & Longitude*

Area Competition Density

The Foursquare API will be used to retrieve information on nearby venues, namely pertaining to the category/type of venue as an indication whether a business is competition. An API call utilizes the latitude, longitude coordinates as well as a predefined RADIUS (among other parameters if necessary) to return a list of venues within the area. The initial result would resemble data similar to the following frame (Figure 2):

	Neighborhood	Venue	Venue Category
0	Parkwoods	Brookbanks Park	Park
1	Parkwoods	GTA Restoration	Fireworks Store
2	Parkwoods	Variety Store	Food & Drink Shop
3	Victoria Village	Victoria Village Arena	Hockey Arena
4	Victoria Village	Tim Hortons	Coffee Shop
5	Victoria Village	Portugril	Portuguese Restaurant

Figure 2 *Formatted dataframe containing venue category data for analysis on neighbourhood competition values*

This data would then be analyzed for the most ***n*** frequent venue categories present in the area with variable ***n*** chosen to maximize relevance of the categories to the result while ignoring unnecessary noise in the data (minimally occurring venue types complicating the analysis). Based on the venue category, in comparison to the target business category (e.g. Coffee Shop/Cafe), a count of the competitors in the area will serve as one component of the overall evaluation metric of whether a business should consider the location.

Toronto BIA Data

The second component of the metric would primarily incorporate the location data provided for each Toronto BIA. The base dataset provides latitude, longitude geometries of the BIAs with central coordinates and identifiers (e.g. AREA_NAME). The distance to each BIA's central point will be determined for each neighbourhood and the BIA will be considered based on the RADIUS as specified during the initial Foursquare venue exploration. The metric would reflect the difference between two areas with similar competition factors based on availability of a BIA within the range of the neighbourhood. To illustrate, if the BIA shown below (Figure 3) fell within the range of one of two prospective neighbourhoods (based on LONGITUDE, LATITUDE as shown) with Coffee Shop-related business competition counts of 10 and 10, the neighbourhood with the BIA in range would rank higher than that of the neighbourhood out of range. Contrarily if both neighbourhoods were within range of the BIA, it would be the competition value that would hold higher weight in regards to their relative ranks.

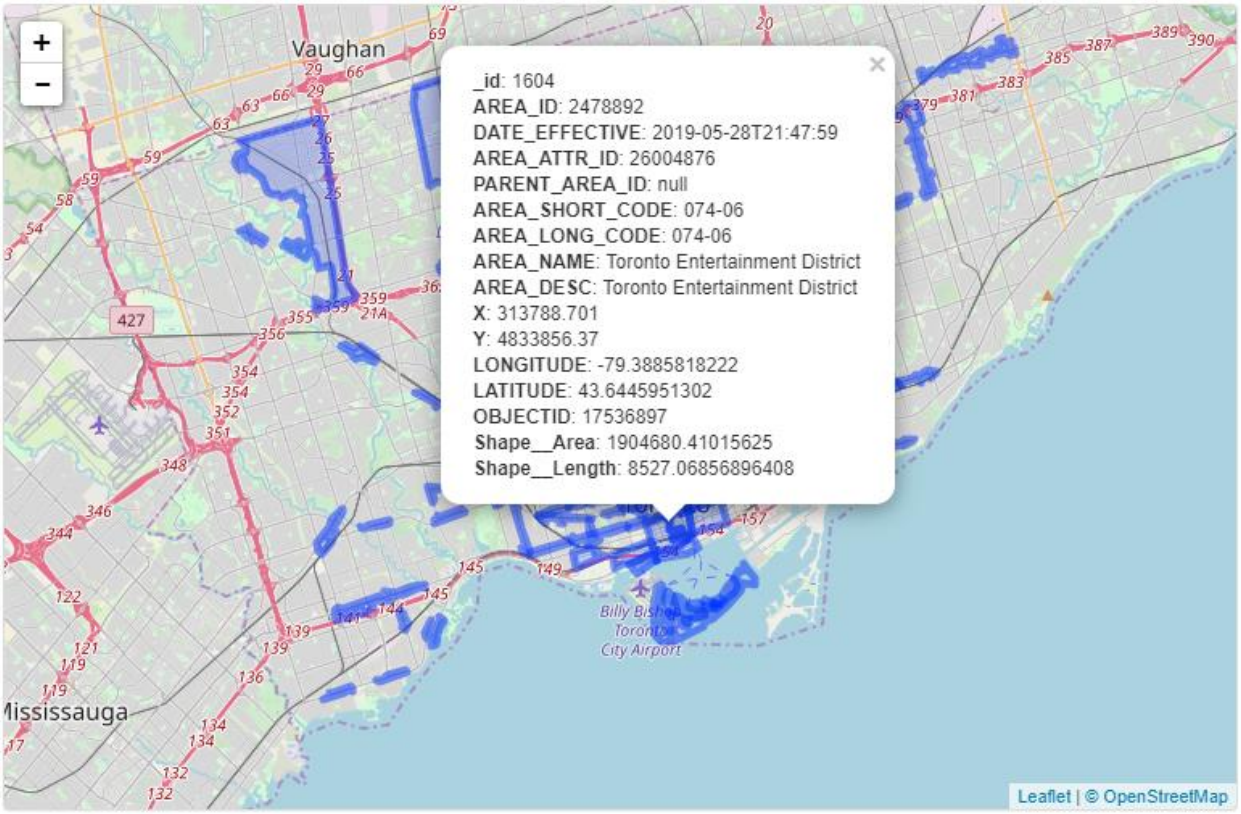

Figure 3 *Data preview of BIA geometry and data features*

METHODOLGY

Besides the data wrangling required to shape the prerequisite base neighbourhood and BIA data, additional analytics are needed to determine the relationship between specific neighbourhoods, corresponding venues and BIA resources before the machine learning analysis can be performed.

Determining Neighbourhood Venue Frequency

As stated prior, one of the main components of establishing the competition density metric is determining the variable ***n*** pertaining to the top *n* frequent venue categories per neighbourhood. The goal during selection of an appropriate value is to ensure that it accounts for the most representative view of venue category distribution for the majority of neighbourhoods. Due to the nature of the Foursquare data venue categories can be have relatively sparse instances. To address this issue $n = 5$ was selected, this provides the ability to gain an insight on individual neighbourhood venue category breakdowns while separating from the instances where "venue category noise" occurs ("venue category noise" corresponds to the numerous venue categories per neighbourhood with counts of 1 location). $n = 5$ corresponds to retrieving the top *5* venue categories for each neighbourhood to incorporate into competition density calculations. For example, a sample of the venue frequency data is shown below, it shows a simple breakdown of the top 5 venue categories per neighbourhood along with the corresponding counts for each category (Figure 4**)**:

	Neighbourhood	1st Most Common Venue	# of 1st Venue	2nd Most Common Venue	# of 2nd Venue	3rd Most Common Venue	# of 3rd Venue	4th Most Common Venue	# of 4th Venue	5th Most Common Venue	# of 5th Venue
0	Adelaide, King, Richmond	Coffee Shop	8	Café	6	Steakhouse	4	Thai Restaurant	3	Cosmetics Shop	3
1	Agincourt	Supermarket	1	Skating Rink	1	Mediterranean Restaurant	1	Seafood Restaurant	1	Shanghai Restaurant	1
2	Agincourt North, L'Amoreaux East, Milliken, St...	Chinese Restaurant	3	Fast Food Restaurant	2	Pizza Place	2	Coffee Shop	1	Malay Restaurant	1
3	Albion Gardens, Beaumond Heights, Humbergate, ...	Grocery Store	3	Pizza Place	2	Hardware Store	1	Auto Garage	1	Coffee Shop	1
4	Alderwood, Long Branch	Convenience Store	2	Pizza Place	2	Pool	1	Donut Shop	1	Sandwich Place	1
5	Bathurst Manor, Downsview North, Wilson Heights	Coffee Shop	2	Pizza Place	2	Gas Station	1	Community Center	1	Sandwich Place	1
6	Bayview Village	Japanese Restaurant	2	Bank	2	Café	1	Shopping Mall	1	Grocery Store	1
7	Bedford Park, Lawrence Manor East	Coffee Shop	3	Italian Restaurant	3	Sandwich Place	2	Fast Food Restaurant	2	Breakfast Spot	1
8	Berczy Park	Coffee Shop	9	Café	5	Beer Bar	4	Restaurant	4	Hotel	4
9	Birch Cliff, Cliffside West	Construction & Landscaping	1	Thai Restaurant	1	College Stadium	1	Diner	1	Café	1

Figure 4 *Dataframe that stores the top 5 venue frequency data per neighbourhood (sample 10 neighbourhoods)*

Exploratory Toronto Neighbourhood & BIA Location Visualization

A form of exploratory visualization was performed to get an idea of the orientation of the neighbourhoods of Toronto relative to the BIAs. Geospatial coordinates for each neighbourhood were used to display location markers superimposed with provided BIA location geometries on the same map. The generated map (Figure 5) establishes a frame of reference for neighbourhood and BIA location densities, namely towards the city's center near the lakeshore. It provides a sense of how the metrics may differ from region to region, the relationship between BIA availability and competition density as you move away from the core of the city.

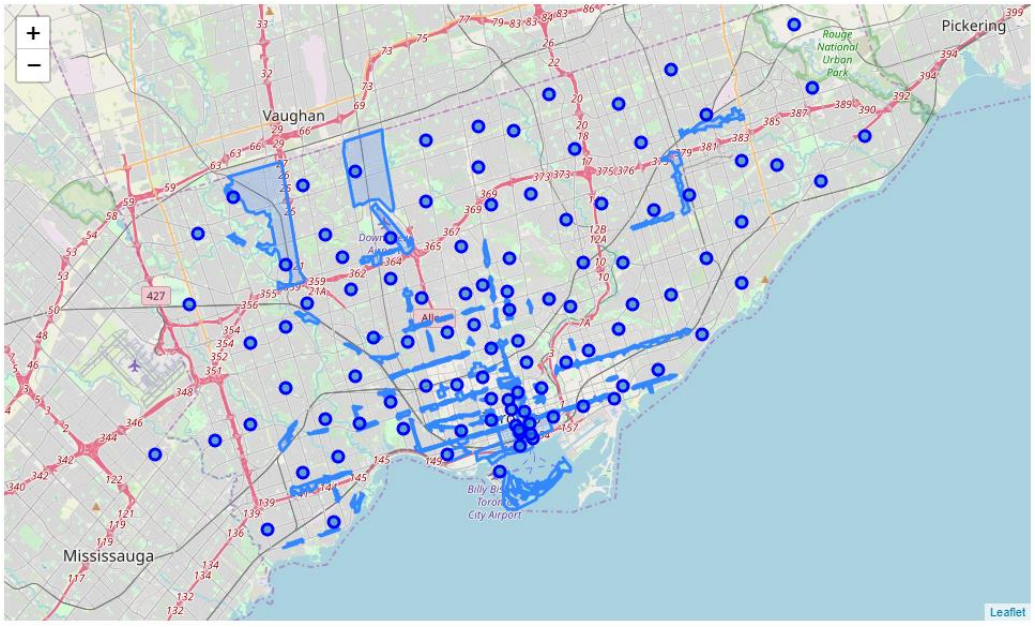

Figure 5 *Toronto neighbourhood & BIA location map visualization*

BIA Availability – Metric Evaluation

In order to determine the *BIA Availability* for each neighbourhood, it's necessary to calculate the distance of each BIA to each neighbourhood location point (using the location data detailed above). Distance calculation is done using a *geodesic* distance function. The geodesic method calculates the distance between two coordinate points as the shortest distance on the surface of an ellipsoidal model of the Earth; it is accurate to round-off and always converges. The default ellipsoid model WGS-84 is used for the calculation due to its acceptance as the most globally accurate.

The computation for a neighbourhood's BIA availability involves evaluating whether the distance of a BIA from the neighbourhood is within the RADIUS as specified during Foursquare venue exploration. Each BIA is checked against each neighbourhood and a count is updated based on whether the distance meets the necessary condition to be considered "available" to the neighbourhood. The results are then consolidated into a dataframe ('Available BIAs' column in Figure 6) which is used to store all the relevant factors for the ML analysis.

Competition Density – Target Business Category Metric Evaluation

Competition Density evaluation leverages the neighbourhood venue frequency data to generate additional features for the analysis. In line with the sample case, the evaluation focuses on the scenario of a prospective entrepreneur looking to open a coffee-related business in the city of Toronto. Based on the possible Foursquare venues, the categories that fit in this particular domain are "Coffee Shop" and "Café" (with the assumption that the venture involves starting a generic coffee-related business, therefore ignoring ethnic and specialty shop categories). The venue frequency data (Figure 4) is then parsed for matching venue categories along with their corresponding counts to add to the factor dataframe (Figure 6). The resulting data forms the basis of features required to perform the project's ML analysis component.

	Neighbourhood	Available BIAs	Coffee Shops	Cafés
0	Adelaide, King, Richmond	2	8	6
1	Agincourt	1	0	0
2	Agincourt North, L'Amoreaux East, Milliken, St...	0	1	0
3	Albion Gardens, Beaumond Heights, Humbergate, ...	0	1	0
4	Alderwood, Long Branch	0	0	0
5	Bathurst Manor, Downsview North, Wilson Heights	0	2	0
6	Bayview Village	0	0	1
7	Bedford Park, Lawrence Manor East	0	3	0
8	Berczy Park	2	9	5
9	Birch Cliff, Cliffside West	0	0	1

Figure 6 *Factors dataframe that stores data features for ML analysis (sample 10 neighbourhoods)*

Competition density is represented as the sum of related business venue categories per neighbourhood. For example, the competition density of the neighbourhood 'Adelaide, King, Richmond' would be *(Coffee Shops + Cafés) = (8 + 6) = 14*. The metric will be explicitly represented when comparing the results of the ML analysis in the following section.

Machine Learning Analysis – *K-Means Clustering*

For this project, *k-means* clustering was performed on the data as the primary ML analysis method. To support the selection of this algorithm amongst others, k-means is used in many data science applications and, with this project, provides the opportunity to discover useful insights from the relevant data in a scenario that requires an unsupervised learning technique. One of the real-world applications of k-means is customer segmentation, a problem domain with characteristics similar to that of the business problem of this project. K-means fits this application as the project involves a form of "*neighbourhood segmentation*", requiring knowledge of neighbourhood BIA availability and competition density serving as each neighbourhood's factors for analysis. In addition, k-means is efficient on the scale of dataset size that this project utilizes (the amount of neighbourhood and BIA data is not large enough to show significant negative effects on performance).

k = 5 was chosen based on observational assessment of the degree of difference between the resulting clusters. The input BIA and competitor data was standardized (using Standard-Scaling) by removing the mean and scaling to unit variance to ensure no discrepancies arise in the results due to the varying scales between BIA numbers and competition counts. These decisions were made to avoid any scenario in which clusters are formed that fail to provide insight on the grounds of improper data processing (i.e. all clusters should have significant member counts with similar characteristics such that meaningful insights can be obtained from the segmentation).

Despite being one of the fastest clustering algorithms available, k-means is susceptible to local minima. One way to address this issue is to run the algorithm multiple times with different centroid initializations. The project was implemented to run the k-means algorithm 20 times with different centroid sees to ensure the final results were the best output of consecutive runs in terms of inertia. Another consideration was made in regards to performance: k-means++ was implemented which selects initial cluster centers using a method that speeds up convergence.

A results dataframe was then created to store the key summarizing factors associated with determining the viability of a particular neighbourhood. A third measure of '*BIA/Comp Ratio*' was established to evaluate the ratio of available BIAs to competitors in the range of a neighbourhood. This measure was used as a high-level comparator between the different clusters to differentiate the level of viability as a new business location. To avoid calculations resulting in *inf*/*NaN* values the ratio was calculated such that all neighbourhoods with a competition density of 0 took the BIA availability as their BIA/Comp ratio. This adjustment allowed areas with low competition counts to be represented in a cluster's mean BIA/Comp ratio statistic.

The following cluster scatter plot visualization (Figure 7) shows a simplistic view of the cluster orientation represented by the relationship between *BIA Availability* and *Competition Density*. The radius of the markers is proportional to the neighbourhood's competition density.

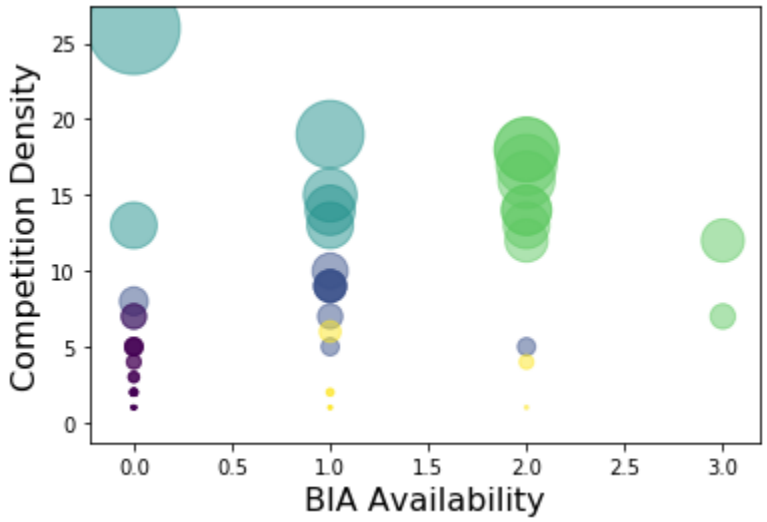

Figure 7 *Cluster scatter plot visualization (Competition Density vs BIA Availability)*

RESULTS

Insights from Cluster Characteristics

Below are views of the members of each cluster from the k-means analysis (cluster 0 – 4, $k = 5$). Above each view is the *mean BIA/Comp ratio* for the cluster (standard deviation provided for reference), which provides a means of comparison between the 5 clusters upon which to make a relative recommendation as to the most viable location for a new coffee-related business. As reference, the *higher* the BIA/Comp ratio, generally the *better* the neighbourhood for a business location in the context of the metrics being used.

Cluster 0 *(results truncated for the report)*

```
Mean BIA/Comp Ratio for Cluster 0 is 0.0
Std Deviation of BIA/Comp Ratio for Cluster 0 is 0.0
```

	Neighbourhood	Available BIAs	Competition Density	BIA/Comp Ratio
2	Agincourt North, L'Amoreaux East, Milliken, St...	0	1	0.0
3	Albion Gardens, Beaumond Heights, Humbergate, ...	0	1	0.0
4	Alderwood, Long Branch	0	0	0.0
5	Bathurst Manor, Downsview North, Wilson Heights	0	2	0.0
6	Bayview Village	0	1	0.0
7	Bedford Park, Lawrence Manor East	0	3	0.0
9	Birch Cliff, Cliffside West	0	1	0.0
10	Bloordale Gardens, Eringate, Markland Wood, Ol...	0	1	0.0
12	Business Reply Mail Processing Centre 969 Eastern	0	0	0.0
14	CN Tower, Bathurst Quay, Island airport, Harbo...	0	0	0.0

Cluster 1

```
Mean BIA/Comp Ratio for Cluster 1 is 0.12873015873015875
Std Deviation of BIA/Comp Ratio for Cluster 1 is 0.11258396454191189
```

	Neighbourhood	Available BIAs	Competition Density	BIA/Comp Ratio
11	Brockton, Exhibition Place, Parkdale Village	1	9	0.111111
15	Cabbagetown, St. James Town	1	9	0.111111
21	Christie	1	5	0.200000
28	Davisville	1	9	0.111111
30	Deer Park, Forest Hill SE, Rathnelly, South Hi...	1	9	0.111111
41	East Toronto	0	8	0.000000
52	High Park, The Junction South	1	7	0.142857
65	Little Portugal, Trinity	2	5	0.400000
69	Northwest	0	7	0.000000
78	Ryerson, Garden District	1	10	0.100000

Cluster 2

```
Mean BIA/Comp Ratio for Cluster 2 is 0.04460831566094723
Std Deviation of BIA/Comp Ratio for Cluster 2 is 0.03548046552664549
```

	Neighbourhood	Available BIAs	Competition Density	BIA/Comp Ratio
19	Central Bay Street	1	13	0.076923
27	Commerce Court, Victoria Hotel	1	19	0.052632
50	Harbourfront	1	15	0.066667
51	Harbourfront East, Toronto Islands, Union Station	1	14	0.071429
74	Rosedale	0	26	0.000000
79	Scarborough Village	0	13	0.000000

Cluster 3

```
Mean BIA/Comp Ratio for Cluster 3 is 0.17496678158442866
Std Deviation of BIA/Comp Ratio for Cluster 3 is 0.09801340874469185
```

	Neighbourhood	Available BIAs	Competition Density	BIA/Comp Ratio
0	Adelaide, King, Richmond	2	14	0.142857
8	Berczy Park	2	14	0.142857
20	Chinatown, Grange Park, Kensington Market	3	7	0.428571
22	Church and Wellesley	3	12	0.250000
32	Design Exchange, Toronto Dominion Centre	2	18	0.111111
44	First Canadian Place, Underground city	2	17	0.117647
49	Harbord, University of Toronto	2	13	0.153846
82	Stn A PO Boxes 25 The Esplanade	2	16	0.125000
83	Studio District	2	18	0.111111
84	The Annex, North Midtown, Yorkville	2	12	0.166667

Cluster 4

```
Mean BIA/Comp Ratio for Cluster 4 is 1.0370370370370372
Std Deviation of BIA/Comp Ratio for Cluster 4 is 0.5098877088847604
```

	Neighbourhood	Available BIAs	Competition Density	BIA/Comp Ratio
1	Agincourt	1	0	1.000000
13	CFB Toronto, Downsview East	1	2	0.500000
16	Caledonia-Fairbanks	1	0	1.000000
31	Del Ray, Keelesdale, Mount Dennis, Silverthorn	1	0	1.000000
47	Glencairn	1	1	1.000000
55	Humber Bay Shores, Mimico South, New Toronto	1	0	1.000000
58	Humewood-Cedarvale	1	0	1.000000
66	Maryvale, Wexford	1	0	1.000000
68	North Toronto West	1	1	1.000000
70	Northwood Park, York University	1	0	1.000000
71	Parkdale, Roncesvalles	2	1	2.000000
73	Queen's Park	1	0	1.000000
77	Runnymede, Swansea	2	0	2.000000
81	St. James Town	2	0	2.000000
85	The Beaches	1	6	0.166667
86	The Beaches West, India Bazaar	1	2	0.500000
87	The Danforth West, Riverdale	2	4	0.500000
93	Weston	1	1	1.000000

The individual cluster breakdown shows that *cluster 4* has the highest mean BIA/Comp ratio with the cluster ranking as follows:

1. **Cluster 4** – Mean BIA/Comp ratio – 1.0370370370370372
2. **Cluster 3** – Mean BIA/Comp ratio – 0.17496678158442866
3. **Cluster 1** – Mean BIA/Comp ratio – 0.12873015873015875
4. **Cluster 2** – Mean BIA/Comp ratio – 0.04460831566094723
5. **Cluster 0** – Mean BIA/Comp ratio – 0.0

By investigation of the cluster members it's apparent that cluster 4 represents neighbourhoods with access to BIAs with minimal to no competition. The locations are all within range of BIA resources and have very few competitors in the neighbourhood, cluster 4 neighbourhoods offer a suitable balance of BIA availability to competition density.

DISCUSSION

Despite the clear standout in the mean BIA/Comp ratio rankings, further interpretation yields additional insights, namely in regards to cluster 0 and cluster 3. Regardless of cluster 4 neighbourhoods providing the best relative all-around environment to start a new coffee-related business, if a particular entrepreneur weighs factors differently, other clusters become more viable. On one hand, if BIA availability were to be desired proportionally more than competition density, cluster 3 neighbourhoods would appear more suitable with greater variety of BIA resources per location. On the other hand, if BIA availability held less significance then cluster 0 neighbourhoods may also (in addition to cluster 4 areas due to their similarity) be worth investigating further as many locations within the cluster show minimal competition. In many ways this analysis is the first step to deciding a new business location, establishing a foundation upon which to assess the general characteristics of Toronto neighbourhoods with different preferences and business goals. Recommendations, at face value, on the grounds of balance between the metrics would be cluster 4 neighbourhoods, although under the right circumstances cluster 0 and cluster 3 neighbourhoods are considerations.

Another factor to consider on an intuitive basis would be the relative locations of the overall clusters to the city's center. One could justifiably make the assertion that neighbourhoods closer to the city's center have higher population densities relative to those further inland. Based on the analysis and mapping performed prior, the general orientation of clusters within the city of Toronto can be visualized (Figure 8 – colour mappings to points given in the caption).

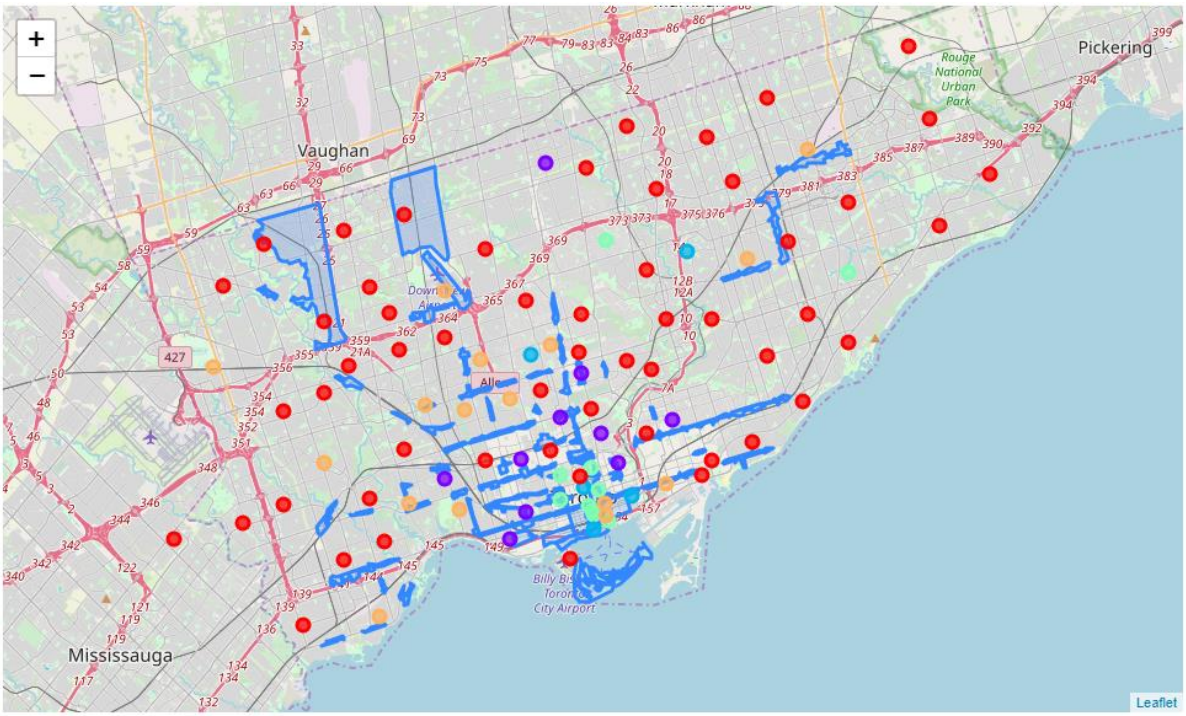

Figure 8 *Toronto neighbourhood clustering visualization*
(red - cluster 0, purple – cluster 1, blue – cluster 2, neon green – cluster 3, orange – cluster 4)

Through investigating the above visualization it's discernable that cluster 4 neighbourhoods occupy a "sweet spot" in terms of their general locations relative to the denser parts of the city. Despite the inferred higher population density in the city's center, one aspect to consider is that increased population density tends to correlate with increased competition density to match the corresponding increase in demand. It's for this reason that cluster 4 occupies a range in the city where it strikes a balance between still being relatively near to the BIA resources of downtown Toronto without being as affected by the greater number of competitors.

Similarly, looking at cluster 3 neighbourhoods it is apparent that the characteristic of moderate-to-high BIA availability to higher competition density is shown as locations are predominantly in the city's center. Cluster 0 neighbourhoods are more widely dispersed and occupy the outskirts of the city, in line with the cluster characteristics of minimal BIA availability coupled with minimal competition density. Geographical cluster positioning could also play a part in recommendations if entrepreneurs prefer locations closer to the denser areas of Toronto. The balanced approach remains to be cluster 4 with clusters 0 and 3 having viability in particular cases.

CONCLUSION

The purpose of this project was to provide entrepreneurs and new business owners the ability to evaluate prospective neighbourhoods in the city of Toronto on the basis of competition and business improvement initiative availability. Foursquare and Toronto Open Data were used to consolidate neighbourhood-level competition density and BIA availability to generate a matric of how viable a particular area was to start a new business (e.g. a coffee-related business in this case). A collection of neighbourhoods with their corresponding measures were then sent through clustering analysis to establish partitions of neighbourhoods with similar characteristics. The k-means clustering analysis yielded neighbourhood segments with relatively unique features which provided a basis for each possible business need. Regardless of the requirements of individual entrepreneurs and business owners, the project provides neighbourhood segmentation that establishes a footing for future selection of areas with varying degrees of BIA availability to competition density (e.g. cluster 4's balanced metrics compared to cluster 3's competition density-leaning neighbourhoods). This project implements for new business owners a resource to evaluate the neighbourhoods of Toronto in a way that considers the business environment, weighing both the availability of BIA initiatives and the level of industry-specific competition.