INITIAL RESULTS FOR POC

Recommendation Solution

Modeling the Identification of Customized Neighborhoods for Real State Purposes

Date: 06/22/2017

Contents

1. Background
2. Methodology
3. Results
4. Discussion
5. Conclusion

1. Background

Background

Why ?

- ✓ **MOTIVATION:** Potential opportunities related to customers who need to relocate to areas where they've never had the chance to live before. Is it possible to pin-point customized homes to those users?

- ✓ **Initial Analysis:** At this moment we will focus on recommending neighborhoods based on a simplified customer profile and a public data set on the city of Toronto, assessing its feasibility. Further analysis are limited at the moment due to the unavailability of free and broken data on market prices and offer of family homes.

Background

Customer perspective: How is life in my future city?

MOVING? COOL!

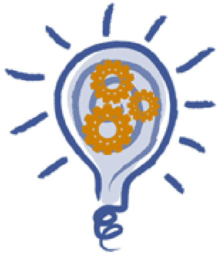
✓ **BUT... :**

- We have never been to Toronto before and, therefore, have no idea about how life in Toronto is.
- How to figure in which area or neighborhood we should look for a home to live?

Potential customers may spend many hours googling information about life in Toronto, about characteristics of some neighborhoods so they may understand which could be the nicest one to live in. Yet, they may end up with little clue about what neighborhood to choose.

Background

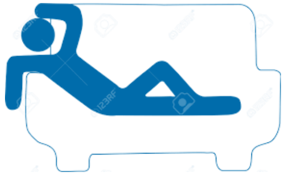
Idea

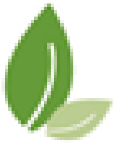


WHAT IF….:

- … there was an easier and faster way to identify a certain neighborhood with a good fit to a certain family profile?

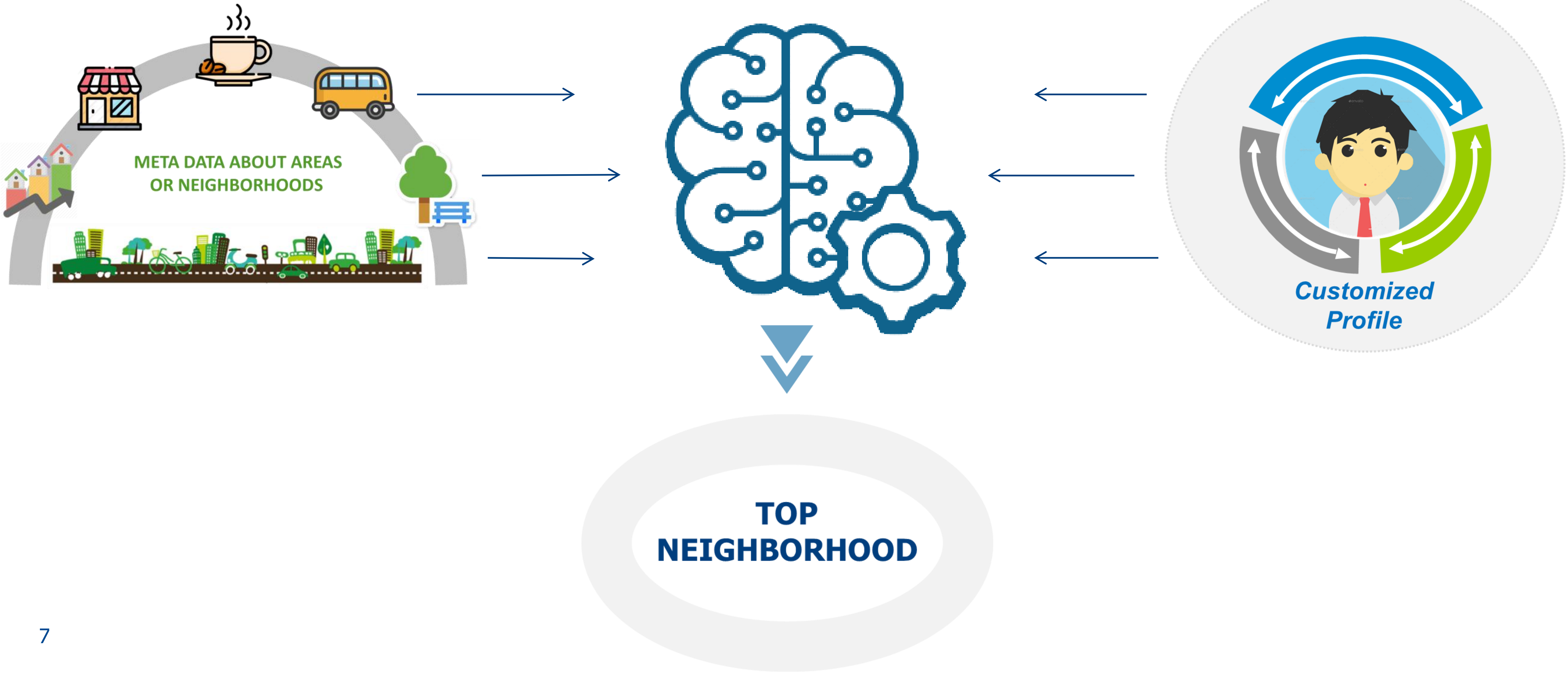
- ➢ Calm or not too noisy?
- ➢ With enough services or conveniences?
- ➢ Green areas or outdoors for leisure?





Background

The Analysis - Fundamentals



META DATA ABOUT AREAS OR NEIGHBORHOODS

Customized Profile

TOP NEIGHBORHOOD

Background

Business Implications

A data analytics approach to such a problem could be used by client …

- ▪ Offer relocation services for families who are moving to another country or city due to any reason;
- ▪ Real estate - customized home or investment opportunities for people who, either due to geographical distance or any other limitations are unable to physically be present in the city/market in question;
- ✓ There could be further benefits from improving client's efficiency and being capable of proposing tailored home locations to their clients, only by having them fill up a form on their profile preferences and allowing the data analytics tool and machine learning algorithms do the rest of the job.

2. Methodology

Methodology

Fundamental Steps

The methodology within the analysis was conducted following the steps listed below:

1. Data Gathering, Handling and Preparation
2. Definition of a Simple Customized User Profile
3. Neighborhood Clustering
4. Cross Evaluation and Cluster Ranking
5. Cross Evaluation – Neighborhood Level Ranking
6. Top Neighborhood(s) - Characteristics Analysis

Methodology

Data Gathering, Handling and Preparation

The data sets considered in the analysis so far are categorized into two distinct groups:

1. Fundamental Data: data considered to be basic and necessary for the development of the analysis

2. Accessory Data: additional data that might be identified as valuable towards the execution of complementary methodologies to further evaluate the results or expand the analysis. The use of Accessory Data might be restrained by the lack of available information or balked by other practical constraints.

Methodology

Data Gathering, Handling and Preparation

The Fundamental Data set to be considered in the project comprises:

- List of Neighborhoods in Toronto, Canada: the data set will be retrieved from Wikipedia at https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M;
- Initial geolocation data on Toronto neighborhoods: this data set will be retrieved from "http://cocl.us/Geospatial_data";
- Further geolocation data on Toronto neighborhoods: these data would include latitude and longitude of each neighborhood in Toronto, Ontario. This data will be collected using geopy package.
- Venue data on Toronto Neighborhoods: general data on venues available for each Toronto neighborhood within a certain radius. These data would include what venues are available, its category, latitude and longitude coordinates, and further venue characteristics, acquired through a foursquare API query.
- A specific customized profile: the profile will include characteristics which will be evaluated as a possible match to a neighborhood.

Methodology

Data Gathering, Handling and Preparation

Accessory Data set that might be considered in this analysis comprises the following:

- Data on real estate market prices for each neighborhood in Toronto, if available.
- Data on population/residents in each Toronto neighborhood.
- Car, transit and traffic data

Most accessory data wasn't available for free. A lot of working hours were oriented to identify data sources for real estate prices per neighborhood in Toronto, information that could be extremely useful to extend the analysis.

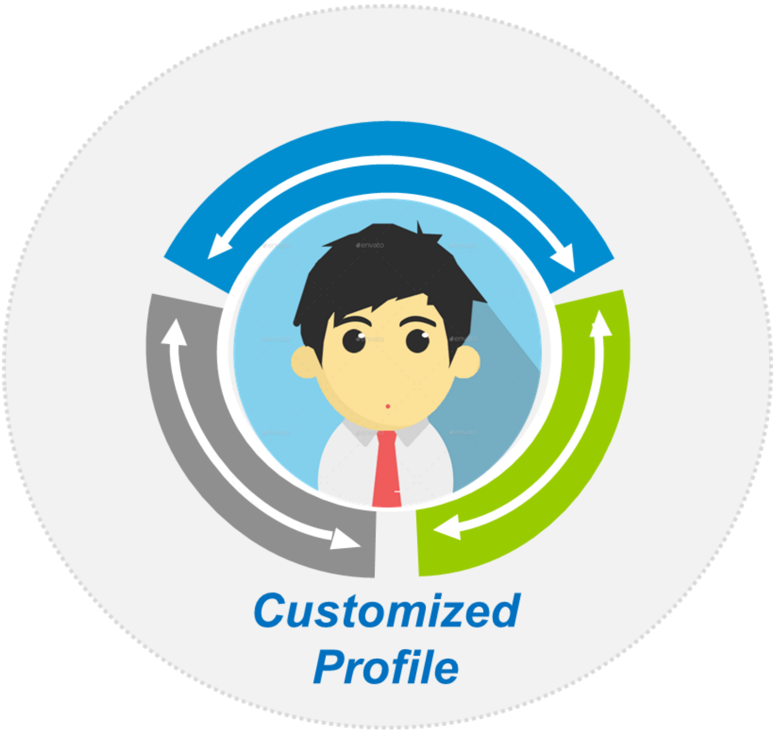
Let's move on!!!!

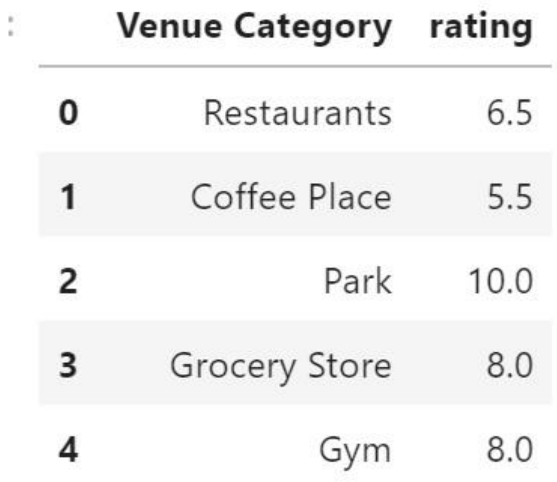
Note: Specific procedures regarding data handling and preparation can be found in the full report at chapter 2, session 2.2.

Methodology

Definition of a simple customized user profile

What sorts of venues are considered more important and with what score (0-10)?





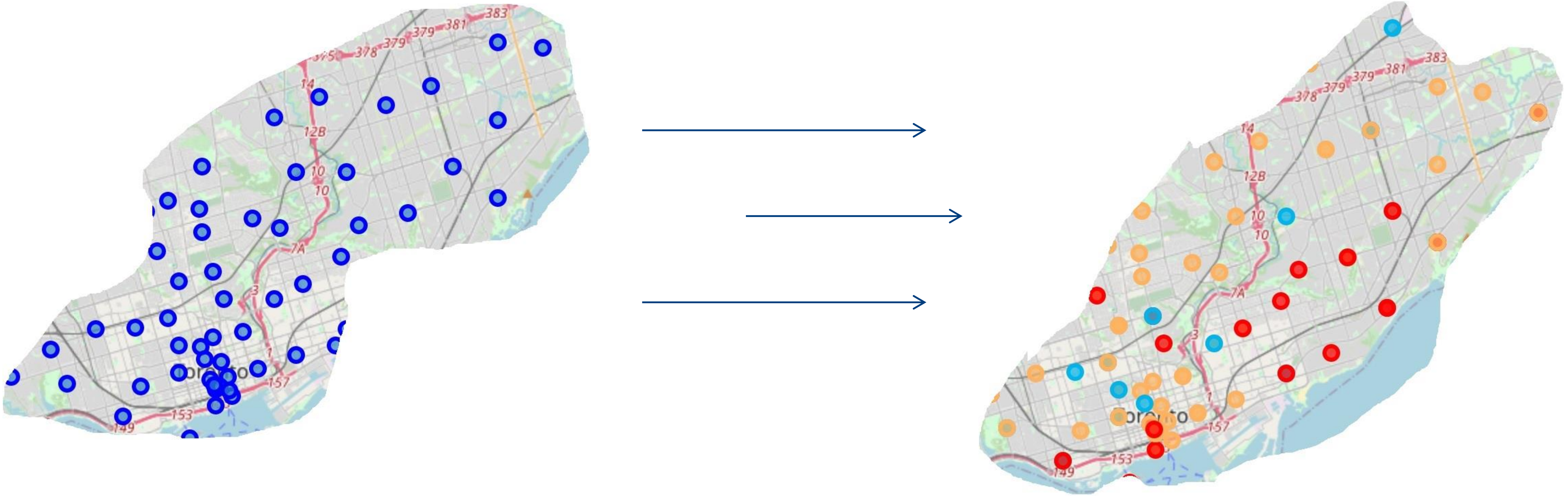
	Venue Category	rating
0	Restaurants	6.5
1	Coffee Place	5.5
2	Park	10.0
3	Grocery Store	8.0
4	Gym	8.0

The future home neighborhood has to offer at least one of each type of venue listed in the profile.

Methodology

Neighborhood Clustering

What neighborhoods share common traces? Can they be clustered into groups with common features?

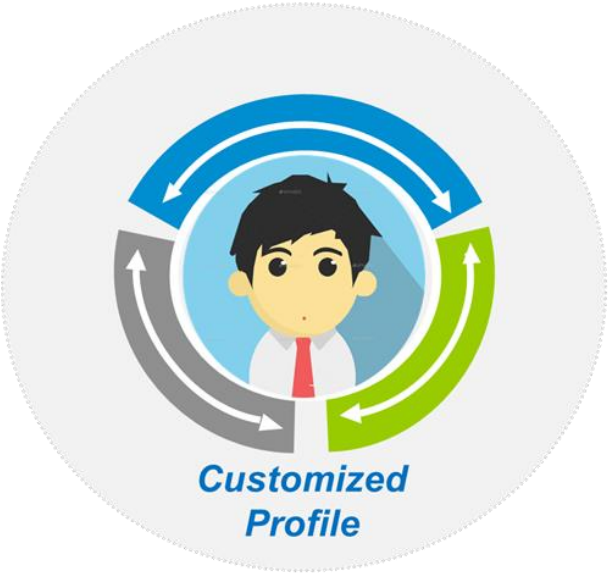


A K-means clustering model was used to group neighborhoods into 5 clusters with random state parameter = 0.

Methodology

Cross Evaluation and Cluster Ranking

A specific KPI was developed and applied to each cluster enabling their ranking based on the type of venues offered and the preferences within the customized profile.



Customized Profile

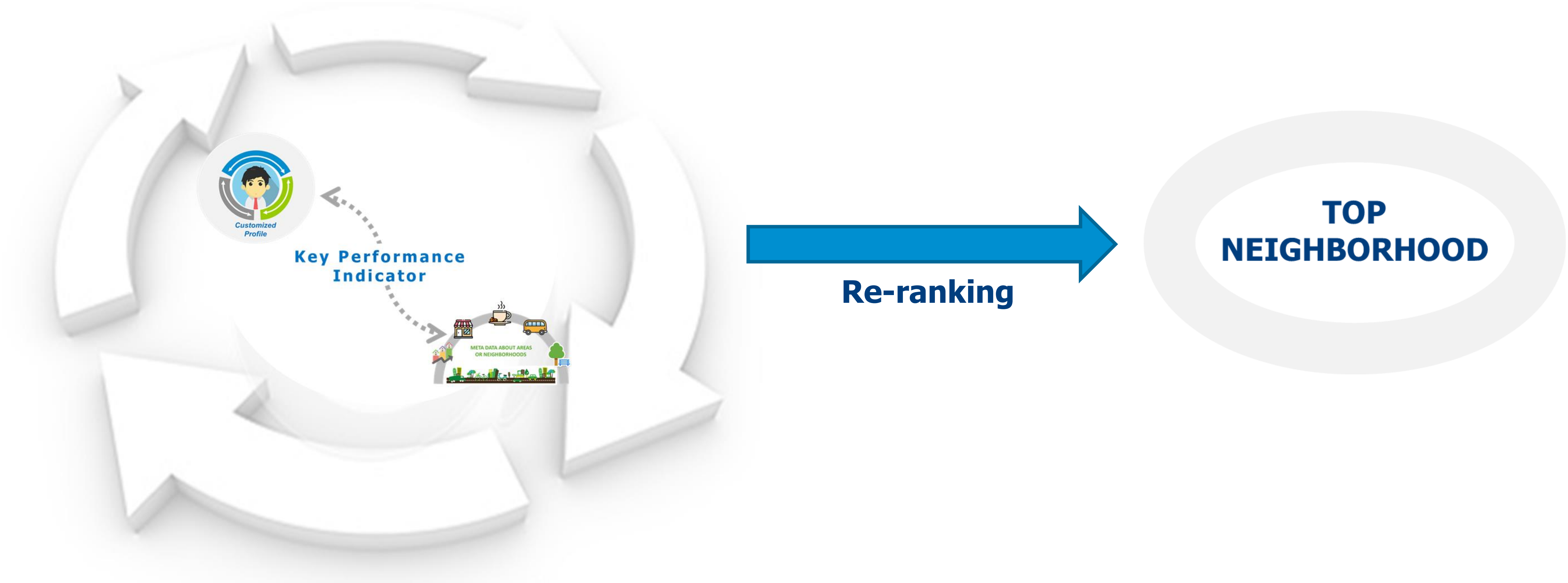
Key Performance Indicator

Note: Specific KPI methodology can be found in the full report at chapter 3, session 3.4.

Methodology

Cross Evaluation – Neighborhood Level Ranking

Once the top cluster is identified, the KPI is applied for every neighborhood within it.



Key Performance Indicator

Re-ranking

TOP NEIGHBORHOOD

Methodology

Top Neighborhood(s) - Characteristics Analysis

What can we tell about the neighborhood(s) identified?

TOP NEIGHBORHOOD

Note: One or more neighborhoods might be identified. This parameter is set by the administrator/user.

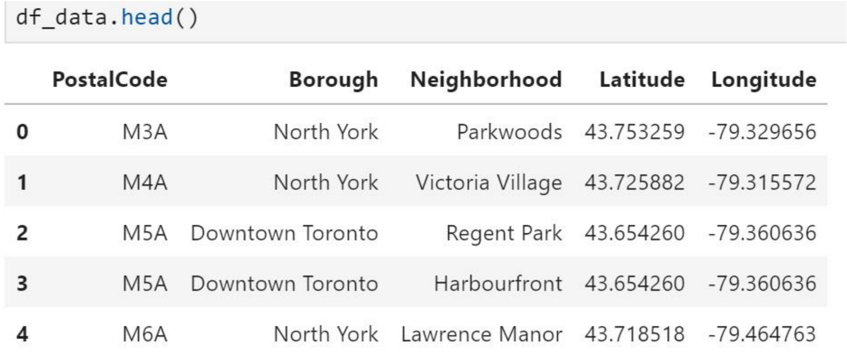
3. Results

Results

Data acquisition, handling, preparation and customer profile definition

The three fundamental data frame objects obtained were the following

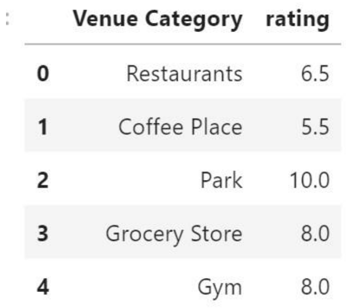
Geospatial Neighborhood Data Set



```
df_data.head()
```

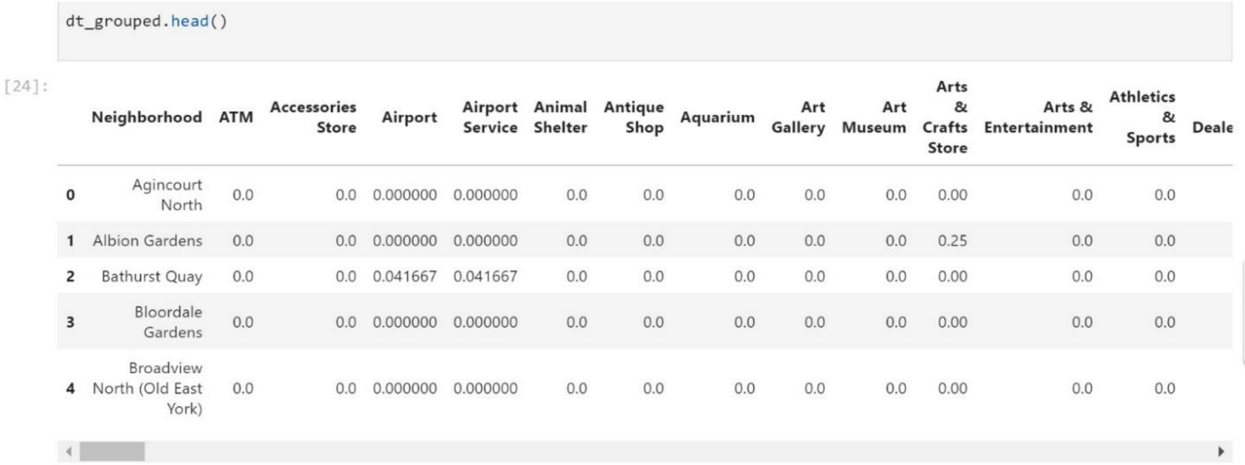
	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park	43.654260	-79.360636
3	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636
4	M6A	North York	Lawrence Manor	43.718518	-79.464763

Customized Profile



	Venue Category	rating
0	Restaurants	6.5
1	Coffee Place	5.5
2	Park	10.0
3	Grocery Store	8.0
4	Gym	8.0

Neighborhood Venue Data Set



```
dt_grouped.head()
```

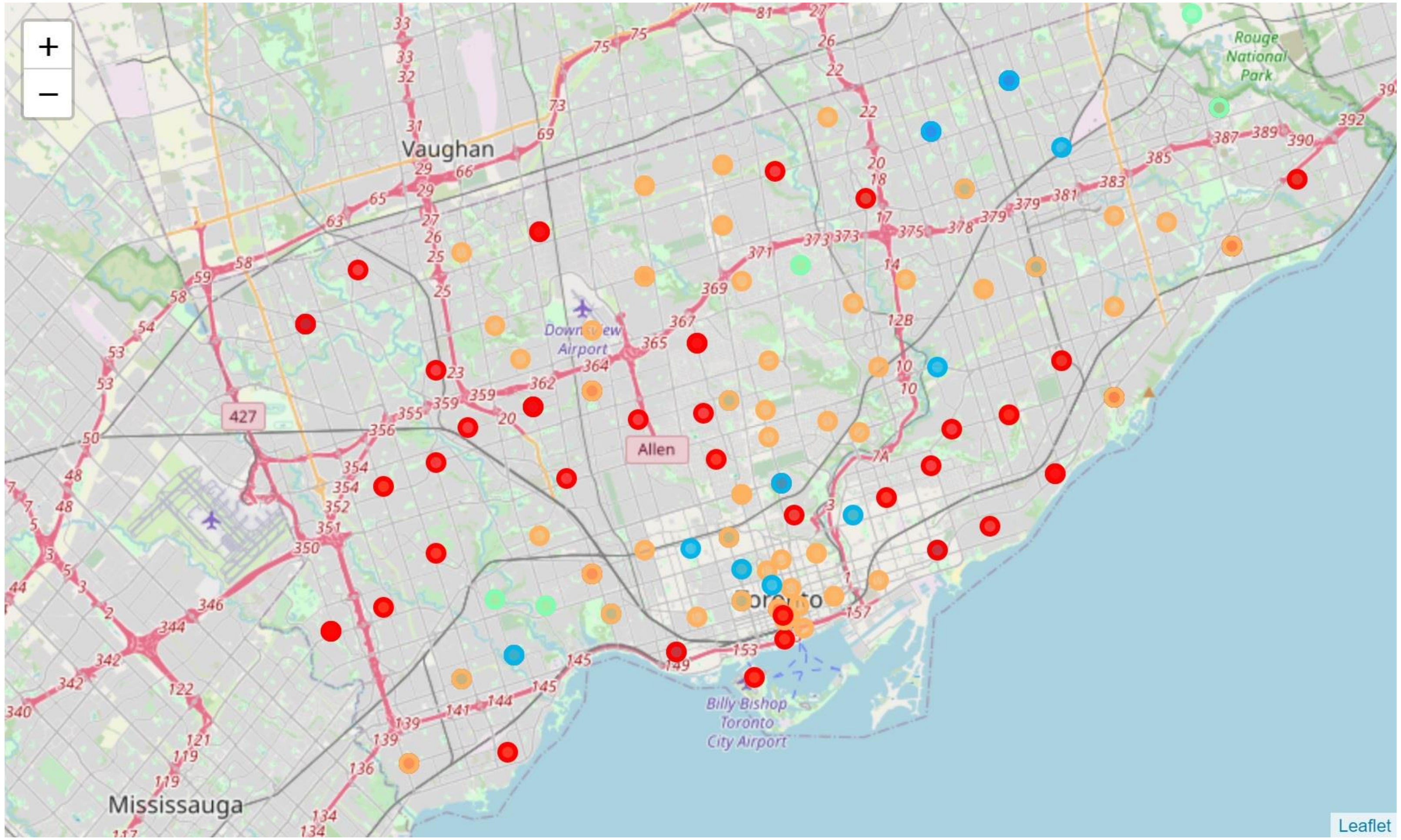
[24]:

	Neighborhood	ATM	Accessories Store	Airport	Airport Service	Animal Shelter	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	Athletics & Sports	Deale
0	Agincourt North	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	
1	Albion Gardens	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.25	0.0	0.0	
2	Bathurst Quay	0.0	0.0	0.041667	0.041667	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	
3	Bloordale Gardens	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	
4	Broadview North (Old East York)	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	

Results

Neighborhood Clustering

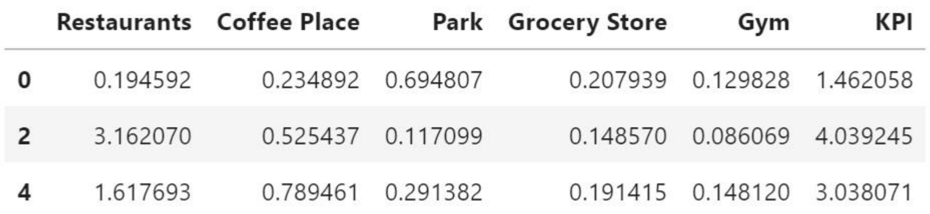
Neighborhoods in Toronto - 5 Clusters pinpointed in the Map



Results

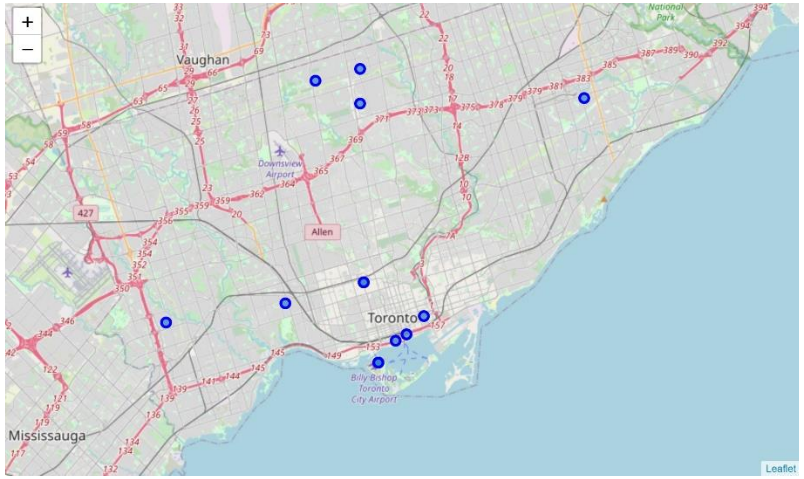
Cross Evaluation and Cluster Ranking

The three clusters had nonzero KPIs



	Restaurants	Coffee Place	Park	Grocery Store	Gym	KPI
0	0.194592	0.234892	0.694807	0.207939	0.129828	1.462058
2	3.162070	0.525437	0.117099	0.148570	0.086069	4.039245
4	1.617693	0.789461	0.291382	0.191415	0.148120	3.038071

Top cluster is number 2 !!!

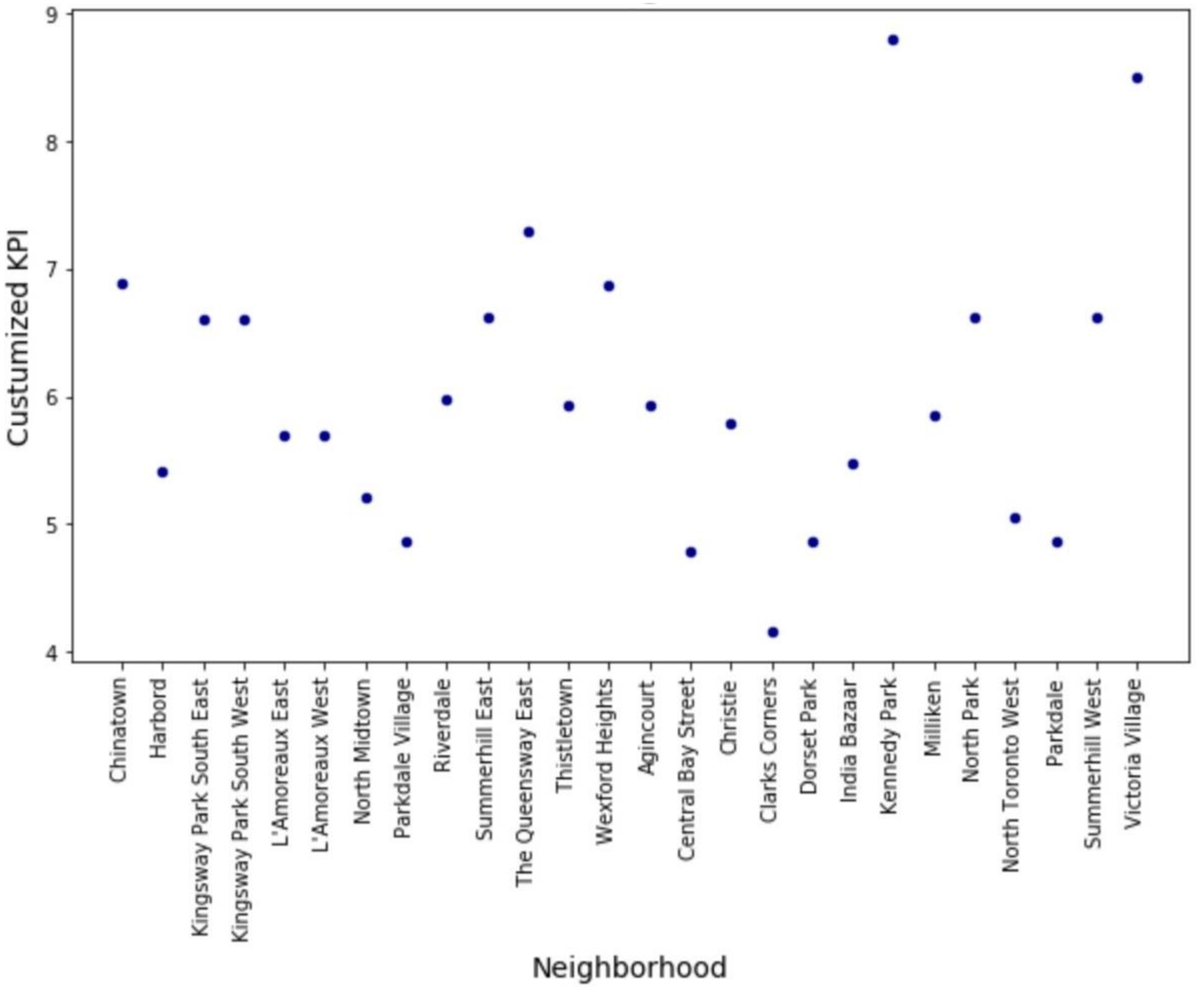


Note: Clusters 1 and 3 either had KPIs equal to zero or did not offer at least one of each type of venue listed in the customized profile. For further details on this assumption check the full report.

Results

Cross Evaluation – Neighborhood Level Ranking

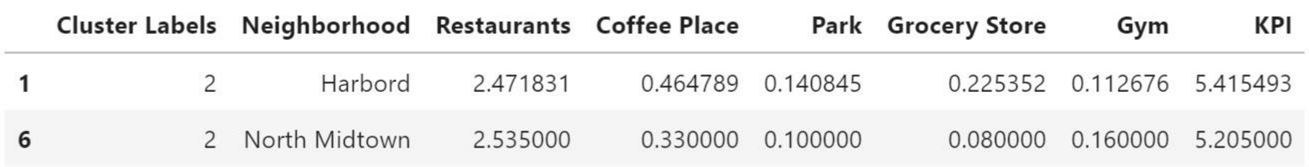
KPIs obtained for each neighborhood within cluster 2



Results

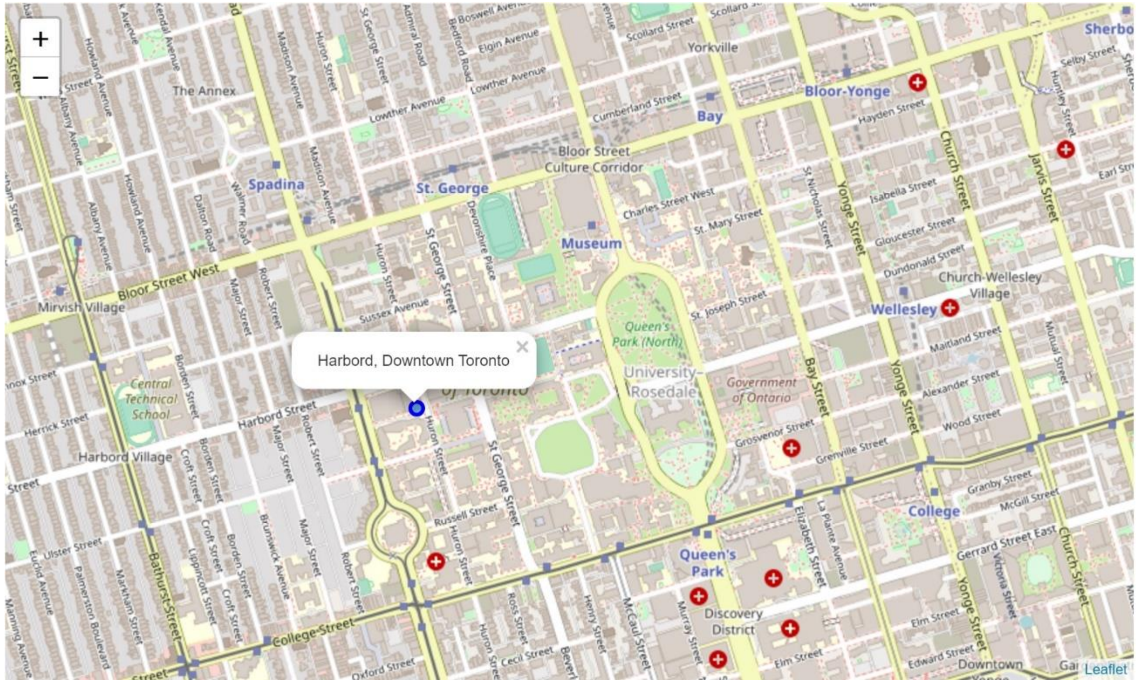
Cross Evaluation – Neighborhood Level Ranking

Ranked neighborhoods within cluster 2



	Cluster Labels	Neighborhood	Restaurants	Coffee Place	Park	Grocery Store	Gym	KPI
1	2	Harbord	2.471831	0.464789	0.140845	0.225352	0.112676	5.415493
6	2	North Midtown	2.535000	0.330000	0.100000	0.080000	0.160000	5.205000

One Top Neighborhood was chosen: *Harbord*

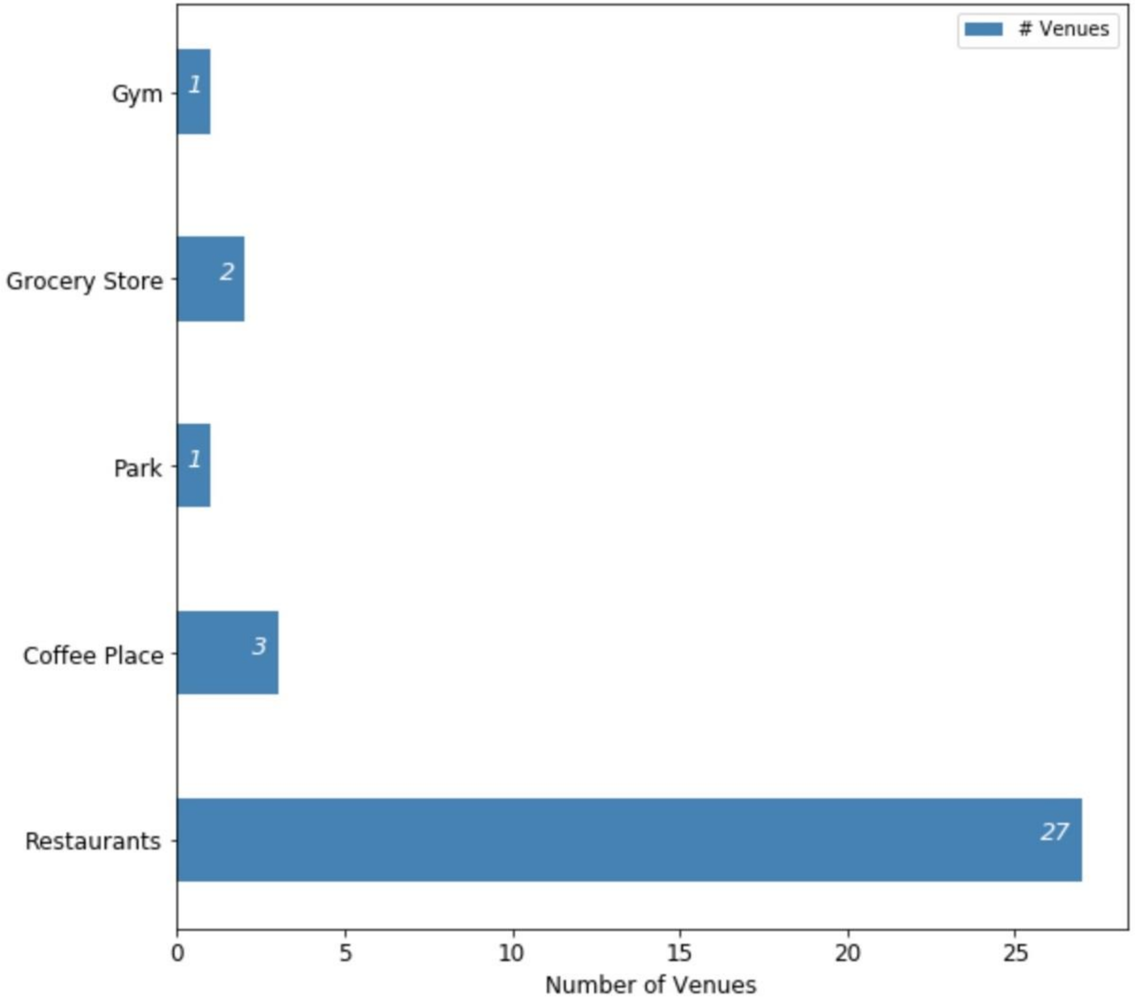


Note: although some neighborhoods achieved high KPIs, only those which had at least one of each venue listed in the customer profile were ranked

Results

Top Neighborhood(s) - Characteristics Analysis

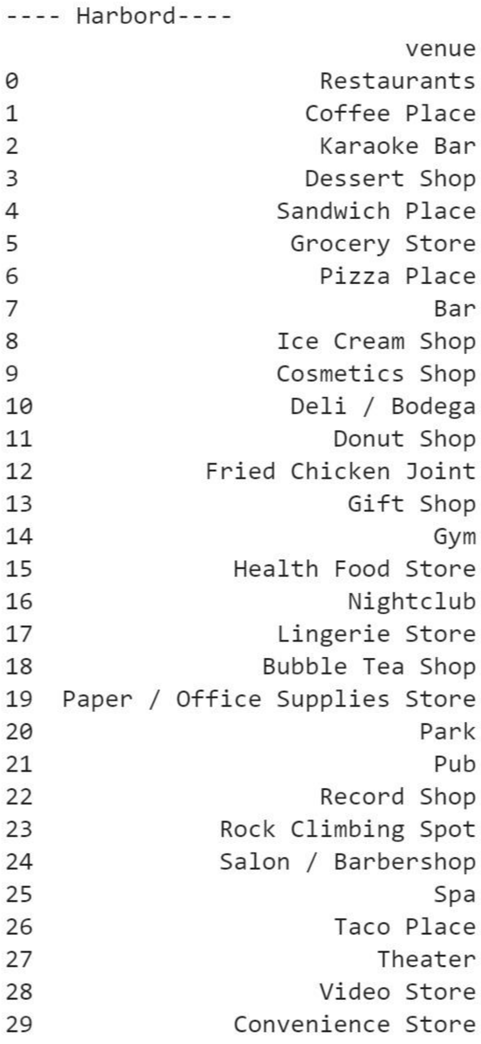
- **How many type of venues Harbord offers as listed in the customer profile?**



Results

Top Neighborhood(s) - Characteristics Analysis

What else?



```
---- Harbord----
                              venue
0                       Restaurants
1                      Coffee Place
2                       Karaoke Bar
3                      Dessert Shop
4                     Sandwich Place
5                     Grocery Store
6                       Pizza Place
7                               Bar
8                     Ice Cream Shop
9                     Cosmetics Shop
10                     Deli / Bodega
11                       Donut Shop
12                Fried Chicken Joint
13                        Gift Shop
14                              Gym
15                 Health Food Store
16                        Nightclub
17                    Lingerie Store
18                   Bubble Tea Shop
19   Paper / Office Supplies Store
20                             Park
21                              Pub
22                      Record Shop
23               Rock Climbing Spot
24               Salon / Barbershop
25                              Spa
26                       Taco Place
27                          Theater
28                      Video Store
29                 Convenience Store
```

Beyond the customized venues, Harbord offers shops, spas, theaters, bakeries and a Rock-Climbing Place

4. Discussion

Discussion

- The described approach could identify two customized neighborhoods based on a simple profile based on scored type of venues
- What if we ignore the clustering algorithm and use a more simplified approach, just like those simple recommender systems?
  - Further analysis on this topic can be found in the full report: results seem too be broad and less accurate.
- K-means clustering seems to have allowed this “recommender system” to better target fit neighborhoods. Although clustering results might be different even under the same dataset, the pinpointed neighborhood is always the same and might only change if the dataset itself is updated. (This was verified under recurrent testing.)
- The model might be upgraded considering further preferences such as budget, level of traffic or population density. The current analysis was limited due to the lack of complementary data available, such as real estate prices data broken down by neighborhood in Toronto.

Conclusion

The Model is efficient in providing customized options of neighborhoods for potential user/companies:

- The number of neighborhoods to be pinpointed is totally customizable, yet, a number too large might lead to inaccurate or misleading results
- The a.i. clustering model bring adaptability to the algorithm and ,yet, high efficiency in its goal of targeting neighborhood(s) with an optimal fit
- An interesting next step would be to refine the model, expanding the preferences within the customer profile and considering further data sets that would allow a cross evaluation on additional preferences (i.e: budget, traffic, population, etc…).