New Service from XYZLifestyle –
Suggested Neighborhoods for New Residents

Kevin Spradlin

August 14, 2021

Table of Contents

Introduction

The COVID-19 epidemic has led to many changes. One of them is an increase in people working remotely from their homes. This change has led many of these types of workers to move to residences in new cities. Once they've moved, people will need to change their existing routines. For example, they will need to find new grocery stores at which to buy their food. Depending on their routines, it might not be easy for them to adapt to their new neighborhoods.

Some people follow customized health plans. These plans describe the types and amounts of food that people should eat and the types and frequency of exercise they should do. A fictional online company, XYZLifestyle.zib, has been creating such plans for their customers for a few years.

Over the past few months, increasing numbers of the company's customers have been asking for a specific type of help. These people have moved and are difficulty finding businesses and other venues in their new neighborhoods that can provide them with the types of food and exercise listed in the health plans they've purchased from the company.

The company wants to provide a new service. It will gather information to determine which neighborhoods in cities have the most health-related venues, such as parks and restaurants with healthier menus. Customers will be able to search through this information, to assist them when they're thinking about moving and want to know which places in a city would be a good fit for them. This new service will help existing customers and may encourage new people to become customers.

To test this new service, a prototype will be created. It will focus on neighborhoods in the city of Toronto.

Data

Wikipedia and the geocoder API will be used to determine the names and locations (latitudes and longitudes) of the various neighborhoods in the prototype's city, respectively. The postal codes of the neighborhoods (from Wikipedia) will be used by geocoder to determine their locations.

Foursquare will be used as the source for data about the venues in the various neighborhoods of the prototype's city. The data lists the venues' names, locations, and categories.

The data from Wikipedia and the geocoder API has been gathered and combined with the data from Foursquare. The category names of the venues will be used to filter the health-related venues from the other ones.

Methodology

A combination of the Wikipedia and geocoder data was used to produce a dataset with locations of 103 neighborhoods in Toronto. Afterwards, geographic map of this data was created to visually confirm that the neighborhoods were in the Toronto area.

Next, this neighborhood/location dataset was combined with venue information from Foursquare to form a larger dataset of various venues in the different neighborhoods of Toronto. The number of neighborhoods in this larger dataset totaled 100, meaning that 3 from the neighborhood/location dataset had no venue data on Foursquare.

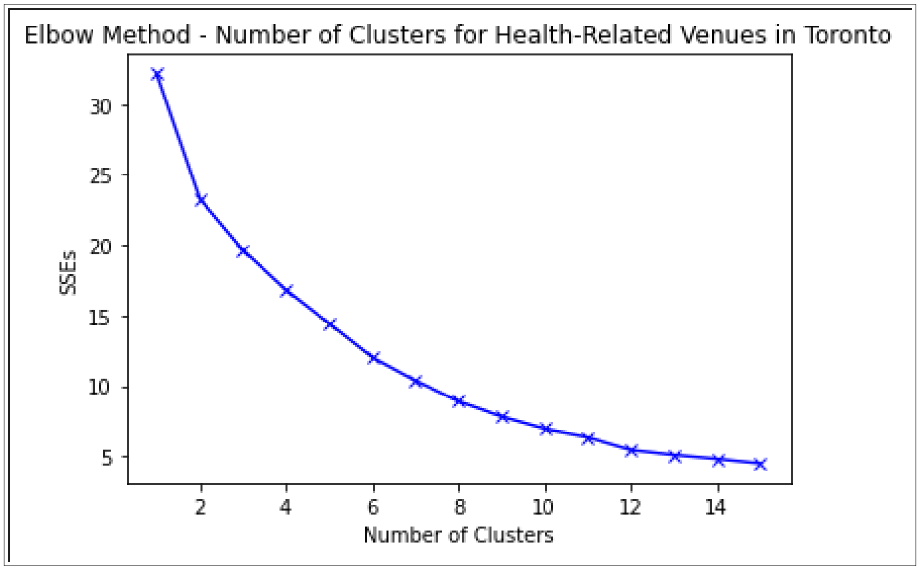
Then, a list of 275 unique venue categories was pulled from Toronto venue dataset. By examining it, 17 were determined to be health-related. Examples include tennis courts, juice bars, and gyms. This list of health-related venues was used to trim down the Toronto venue dataset into one focused on health-related venues. This smaller dataset has data from 68 neighborhoods.

This smaller dataset was converted into a new one that could be used with k-means cluster modeling. The new one has frequencies of the heath-related venues in each neighborhood. In order to confirm that frequency-creation process worked correctly, the five most common heath-related venues in each of the neighborhoods was printed and examined. Only health related venues were found and all of the frequencies ranged from 0.0 to 1.0.

Following the validation of the dataset, a k-means cluster model was used to consolidate the information about the neighborhoods with health-related venues. A simple approach of providing the company's customers with a list of neighborhoods with any type of health-related venues wouldn't be very helpful. For example, some customers swim in order to stay healthy. They would want to know which neighborhoods have pools, as opposed to knowing the neighborhoods that have any type of venue.

With a k-means cluster model, the neighborhoods could be grouped by the most frequent types of venues found in each one. The customers could then learn which groups of neighborhoods have the types of venues that most interest them. Having more granular information about the health-related venues would provide a more valuable benefit to the customers.

In order to determine how many clusters to use in the model, the elbow method was used. 1 to 16 clusters were used to fit a k-means model and the sum of squared errors of each model was recorded. This information, number of clusters versus the model error, was put into a chart (see below).



There is an "elbow" when 2 clusters are used, but it isn't definitive. The model error continues to drop by significant amounts as more clusters are used. A different approach was used to determine the number of clusters. I started with 12 of them and then incrementally reduced them until there weren't any clusters with overlapping types of venues. For example, when 12 clusters was used to fit the model, 2 of them were associated with parks. II found that if 8 clusters were used to fit the model, then no clusters had any overlapping types of venues.

A final k-means model was fit using the health-related venue dataset and 8 clusters. This information was placed onto a map to determine if any location-associated patterns could be seen with the clusters – none were found. Lastly, the neighborhoods and most common health-related venues of each cluster were printed, in order to determine the most-frequent venues in them.

Results

Eight clusters were formed, in which 68 neighborhoods with health-related venues were grouped. Table 1 has summary information on the clusters. Detailed information about the clusters themselves is described in the remaining part of this section.

Table 1 – Summary information on the neighborhood clusters

Most Common Venue(s) in a Cluster	Number of Neighborhoods in Cluster
Juice bars	3
Trails	5
Athletic and sport venues	5
Parks, yoga studios	12
Pools, baseball fields	4
Gyms	6
Parks	15
No common venues	18

Three groups of neighborhoods formed a cluster because juice bars were the most common venue in them, followed by parks and health food stores. They are listed in Table 2.

Table 2 – Neighborhoods associated with juice bars

Neighborhood(s)	Postal Code
Fairview, Henry Farm, Oriole	M2J
Bedford Park, Lawrence Manor East	M5M
Willowdale South	M2N

Five groups of neighborhoods formed a cluster because trails were the most common venue in them, followed by parks and athletic and sport venues. They are listed in Table 3.

Table 3 – Neighborhoods associated with trails

Neighborhood(s)	Postal Code
Humewood-Cedarvale	M6C
The Beaches	M4E
North Park, Maple Leaf Park, Upwood Park	M6L
Forest Hill North & West	M5P
Moore Park, Summerhill East	M4T

Five groups of neighborhoods formed a cluster because athletic and sport venues (e.g. - The Hangar @ Downsview Park) were the most common venue in them, followed by parks, gyms, and fitness centers. They are listed in Table 4.

Table 4 – Neighborhoods associated with athletic and sport venues

Neighborhood(s)	Postal Code
Parkview Hill, Woodbine Gardens	M4B
Woodbine Heights	M4C
Cedarbrae	M1H
Downsview Northwest	M3N
Alderwood, Long Branch	M8W

Twelve groups of neighborhoods formed a cluster because parks and yoga studios were the most common venues in them, followed by gyms, fitness centers, and athletic and sport venues. They are listed in Table 5.

Table 5 – Neighborhoods associated with parks and yoga studios

Neighborhood(s)	Postal Code
Regent Park, Harbourfront	M5A
Garden District, Ryerson	M5B
Bathurst Manor, Wilson Heights, Downsview North	M3H
Thorncliffe Park	M4H
Dufferin, Dovercourt Village	M6H
Golden Mile, Clairlea, Oakridge	M1L
Studio District	M4M
Davisville North	M4P
North Toronto West	M4R
St. James Town, Cabbagetown	M4X
The Kingsway, Montgomery Road, Old Mill North	M8X
Enclave of M4L	M7Y

Four groups of neighborhoods formed a cluster because pools and baseball fields were the most common venues in them, followed by parks and athletic and sport venues. They are listed in Table 6.

Table 6 – Neighborhoods associated with pools and baseball fields

Neighborhood(s)	Postal Code
Hillcrest Village	M2H
Downsview Central	M3M
Humberlea, Emery	M9M
Old Mill South, King's Mill Park, Sunnylea	M8Y

Six groups of neighborhoods formed a cluster because gyms were the most common venue in them, followed by athletic and sport venues and parks. They are listed in Table 7.

Table 7 – Neighborhoods associated with gyms

Neighborhood(s)	Postal Code
Don Mills North	M3B
Don Mills South	M3C
Humber Summit	M9L
Enclave of L4W	M7R
New Toronto, Mimico South, Humber Bay Shores	M8V
Mimico NW, The Queensway West	M8Z

Fifteen groups of neighborhoods formed a cluster because parks were the most common venue in them, followed by athletic and sport venues, gyms, and vegetarian and vegan restaurants. They are listed in Table 8 (note that this cluster differs from the one described in Table 5 in that this cluster's most common venue in all neighborhoods is parks while the most common venue in Table 5 was parks in some neighborhoods and was yoga studios in others).

Table 8 – Neighborhoods associated with parks

Neighborhood(s)	Postal Code
Parkwoods	M3A
Caledonia-Fairbanks	M6E
Christie	M6G
The Danforth East	M4J
Downsview East	M3K
Downsview West	M3L
India Bazaar, The Beaches West	M4L
Willowdale, Newtonbrook	M2M
Lawrence Park	M4N
York Mills West	M2P
High Park, The Junction South	M6P
The Annex, North Midtown, Yorkville	M5R
Kingsview Village, St. Phillips, Martin Grove	M9R
Milliken, Agincourt North, Steeles East	M1V
Rosedale	M4W

Eighteen groups of neighborhoods formed what appears to be a residual cluster, containing the neighborhoods that weren't close enough to the ones in any of the other clusters. There isn't a single most common venue. The more common ones in this residual cluster range from gyms to yoga studios to salad places. The neighborhoods and postal codes are listed in Table 9.

Table 9 – Neighborhoods in the residual cluster

Neighborhood(s)	Postal Code
Ontario Provincial Government	M7A
St. James Town	M5C
Berczy Park	M5E
Central Bay Street	M5G
Richmond, Adelaide, King	M5H
Harbourfront East, Union Station, Toronto Islands	M5J
Little Portugal, Trinity	M6J
The Danforth West, Riverdale	M4K
Toronto Dominion Centre, Design Exchange	M5K
Brockton, Parkdale Village, Exhibition Place	M6K
Commerce Court, Victoria Hotel	M5L
Davisville	M4S
University of Toronto, Harbord	M5S
Runnymede, Swansea	M6S
Kensington Market, Chinatown, Grange Park	M5T
Enclave of M5E	M5W
First Canadian Place, Underground city	M5X
Church and Wellesley	M4Y

Discussion

The neighborhood clusters produced by the methodology have useful information. Customers can provide neighborhoods with residences they're considering to buy or rent, and the company can tell them about the most common types of health-related venues that they can find in those neighborhoods.

The approach of using k-means cluster modeling to consolidate the information of the neighborhoods with health-related venues was useful but also has some drawbacks. Many clusters contain neighborhoods with very similar rankings of health-related venues. For example, the most common venue in the neighborhoods in the cluster described in Table 2 is the same for each neighborhood: juice bars. Any customers who are interested in this particular type of health-related venue can learn the neighborhoods in Toronto that most frequently contain those venues, because of the use of cluster modeling. On the other hand, some clusters didn't have any easily-identified common venues. For example, the most common venues in the neighborhoods in the cluster described in Table 5 were: parks, gyms, yoga studios, and athletic and sport venues.

Another concern is that some types of health-related venues aren't visible with a cluster modeling approach. For example, if a customer was interested in tennis courts, then the cluster modeling approach couldn't provide them with helpful information. The approach didn't identify any group of neighborhoods in which tennis courts were the most common type of health-related venue.

Instead of using clusters, the company could simply show customers the health-related venues in the neighborhoods the customers provide. Another option would be for the customers to select one or more health-related venues and the company would provide them with a list of neighborhoods with those venues. Alternatively, a different clustering algorithm could be used with the data, like hierarchical or density-based clustering.

Due to relative ease of gathering the data, it would be useful to determine well this project can work when it examines other major cities, like New York or Los Angeles.

Lastly, regardless of the approach that's used to identify the neighborhoods with health-related venues, this data will have to be gathered or refreshed with some regular frequency. Over time, new venues can open while old venues can close or relocate. Consequently, the number and types of health-related venues in each of Toronto's neighborhoods can and will change over time.

Conclusion

Useful information on neighborhoods with health-related venues in Toronto was uncovered by combining neighborhood, location, and venue data from three sources. Using a k-means cluster modeling approach to summarize that data led to a product that can help some of the company's existing and future customers. It's likely that one of the findings from this prototype project will be that a sizable number of customers won't gain anything. If that occurs, then new approaches to summarizing the data should be tested in future projects.