Finding a Neighbourhood to Call Home

Data Science Capstone Project Presentation
Reid J.C., 2020-06-30

Table of Contents:

---

1. Introduction

1.1 Background

A young couple have finished school and have been saving money while living in a small apartment. The couple is looking to move to the western edge of Lake Ontario, because they know they don't want to live in the heart of Toronto, and they also have family in the GTHA (Greater Toronto and Hamilton Area). They're looking into buying a house but are overwhelmed due to the significant number of important factors required to narrow down the regions that they should invest their time and energy into to explore further. As a young couple, they are interested in the notion of future kids being able to walk to school, as well as nearby daycare centres being available since both adults are employed. The couple is also highly interested in being able to enjoy their local neighbourhood's outdoor and recreational activities, but they find it difficult to explore this in a time-efficient manner and have so far been overwhelmed looking at city directories and Google Maps.

1.2 Problem Description

Purchasing a house is a significant investment, and there are many big decisions to make and many unknowns to uncover prior to signing a mortgage.

- What city should they live in?
- What is the composition of each city?
- Where are the neighbourhoods within each city?
- What types of neighbourhoods offer the amenities they are interested in?
- Which neighbourhoods are similar to each other?

The couple is looking for a potential neighbourhood to move into that will continue to provide local activities for them and potential future kids and looking into neighbourhoods within west Hamilton, Burlington, Oakville, and Mississauga. The couple want daycare and elementary schools to be local, as well as a high diversity of outdoor and athletic facilities. The couple have a lot of outdoor facilities and athletic facilities in mind that they are interested in, including:

- Playground
- Bike Trail
- Beach
- Botanical Garden
- Dog Run
- Fishing Spot

- Lake
- Park
- Pool
- Boxing Gym
- Climbing Gym
- Cycle Studio
- Gyms
- Martial Arts Dojo
- Outdoor Gym
- Pilates Studio
- Track
- Yoga Studio
- Indoor Play Area
- Recreation Center

A search of this magnitude (e.g. 4 cities and many neighbourhoods within them, across so many different types of amenities, and factoring in geolocation) would take up an extraordinary amount of time to individually search per neighbourhood and manually comb the results (i.e. Google searches). Therefore, the couple are looking for a data science approach that could help them pinpoint the neighbourhoods that offer the amenities they are interested in, as well as the neighbourhoods based on the diversity and quantity of amenities offered. Furthermore, the couple is interested in knowing generally which neighbourhoods are more and less similar to each other based on these amenities.

1.3 Target Audience

This problem is relevant to other couples and young adults in similar stages of life, who for personal reasons are interested in the communities to the west of Toronto, Ontario, Canada. However, this type of analysis could be adapted by anyone looking to find a desirable neighbourhood, by adapting he search criteria to suit their own personal preferences.

---

2. Data

2.1 Neighbourhood names and geolocations

Unfortunately, the Wikipedia page for the neighbourhood and postal code information for this region is incompletely annotated (Wikipedia URL). However, the following database (Canada Postal Codes), is available to curate this information for the following cities: Hamilton (west), Burlington, Oakville, and Mississauga. Conveniently, neighbourhoods are grouped by postal code so as to have roughly even neighbourhood areas. Two small postal code areas were joined (i.e. Burlington Central and Maple regions) into one neighbourhood, since they are both relatively small territories. The neighbourhood of Malton was excluded from Mississauga due to the highly urban location of this neighbourhood and proximity to the airport. After preprocessing the data using the aforementioned descriptions, 35 distinct neighbourhoods were assembled and used for subsequent analysis.

	City	Neighbourhood	Postal_code	Latitude	Longitude
0	Mississauga	Applewood, Dixie	L4Y	43.602800	-79.592900
1	Mississauga	East Hurontario, West Rathwood	L4Z	43.619200	-79.653800
2	Mississauga	Mississauga Valley, East Cooksville	L5A	43.588300	-79.609100
3	Mississauga	City Centre, Cooksville, Fairview, East Credit...	L5B	43.577100	-79.630600
4	Mississauga	West Creditview, Mavis, Erindale	L5C	43.562400	-79.650400
5	Mississauga	Lakeview	L5E	43.583600	-79.561000
6	Mississauga	SW Lakeview, Mineola, East Port Credit	L5G	43.564700	-79.585200
7	Mississauga	Lorne Park, West Port Credit	L5H	43.541900	-79.616400
8	Mississauga	Park Royal, Clarkson, Birchwood, Rattray Park ...	L5J	43.510200	-79.629600
9	Mississauga	Sheridan Homelands, Sherwood Forest	L5K	43.527200	-79.661700
10	Mississauga	Erin Mills, Business Park	L5L	43.537300	-79.690300
11	Mississauga	North Erin Mills, Churchill Meadows, Streetsvi...	L5M	43.563700	-79.720200
12	Mississauga	Lisgar, Meadowvale West	L5N	43.592400	-79.761100
13	Mississauga	West Hurontario	L5R	43.606000	-79.670800
14	Mississauga	East Credit	L5V	43.597200	-79.693100
15	Mississauga	Meadowvale	L5W	43.631300	-79.714800
16	Oakville	N: Wedgewood Creek, Winston Park, River Oaks, ...	L6H	43.480397	-79.702388
17	Oakville	E: Ford Drive, Morrison, Clearview, Old Oakville	L6J	43.475700	-79.659100
18	Oakville	Central Oakville	L6K	43.439600	-79.687800
19	Oakville	S: Bronte, West Oakville, SW Oakville	L6L	43.412540	-79.710121
20	Oakville	NW: West Mount, West Oak Trails, Glen Abbey	L6M	43.445136	-79.744617
21	Burlington	NE: Elizabeth Gardens, Longmoor, Pinedale, Sho...	L7L	43.369117	-79.756033
22	Burlington	N: Alton Village, Headon Forest, Palmer, Millc...	L7M	43.382747	-79.808231
23	Burlington	E: Roseland, Dynes	L7N	43.349176	-79.782273
24	Burlington	W: Tyandaga, Mountainside	L7P	43.355300	-79.841400
25	Burlington	S: Central, Plains, Maple	L7R, L7S	43.332300	-79.807732
26	Burlington	SW: Aldershot, La Salle	L7T	43.308100	-79.850700
27	West Hamilton	Durand, Kirkendall, Chedoke Park	L8P	43.251400	-79.892500
28	West Hamilton	Central, Strathcona, South Dundurn	L8R	43.272600	-79.879200
29	West Hamilton	Westdale, Cootes Paradise, Ainslie Wood	L8S	43.260600	-79.916100
30	West Hamilton	East Mountain	L9A	43.227100	-79.871000
31	West Hamilton	West Mountain	L9C	43.231300	-79.904900
32	West Hamilton	Dundas	L9H	43.263800	-79.952400
33	West Hamilton	Ancaster	L9K	43.216840	-79.985600
34	West Hamilton	Waterdown	L0R	43.325531	-79.901658

2.2 Foursquare data

Foursquare offers customizable searches based on a particular type of venue, which is codified using a category ID. The category ID's for various establishments are available from Foursquare (URL).

2.3 The Couple's Neighbourhood Criteria:

- Must have local daycare or childcare service
- Must have local elementary school
- Local fitness amenities, i.e. gyms, pools, dojo, studios; and outdoor amenities, i.e. trails, parks, dog run, fishing, playground, etc.

Therefore, sequential calls to the Foursquare API will be used to generate the data for the following:

- day care: 5744ccdfe4b0c0459246b4c7
- elementary school: 4f4533804b9074f6e4fb0105
- outdoors and recreation: 4d4b7105d754a06377d81259

The data will be manipulated using pandas and numpy packages in python. Neighbourhoods will be removed if they do not have a day care or elementary school within 1000 metres of their neighbourhood's geolocation, which will be confirmed using map visualization packages (folium) to represent the middle of the centre of the neighbourhood. Then, neighbourhoods will be evaluated for outdoor and recreation amenities. They will be ranked by both quantity and diversity, and clustered (K Means clustering) to evaluate neighbourhoods that are similar in the types of outdoor and recreation activities they offer. Map visualization packages will be used to identify similar neighbourhoods, as well as the highly desirable neighbourhoods based on the couple's criteria.

---

3. Methodology

3.1 Analytical Approach

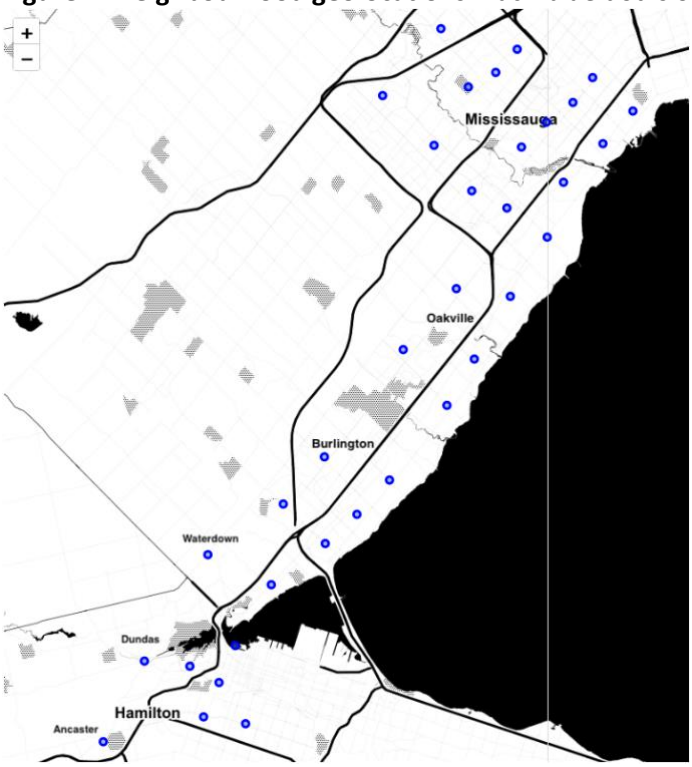
Here I describe the main components of the report and the exploratory data analysis performed, including inferential statistical testing and machine learning techniques used. The initial exploratory analysis is called from the Foursquare API as described below. Briefly, the neighbourhoods are evaluated for the presence of daycare centres, and then elementary schools. If a neighbourhood does not have one or the other amenity, is removed. Then, a wide array of recreational and outdoor facilities were investigated for the remaining neighbourhoods.

Principal component analysis (PCA) was used to cluster the neighbourhoods based on recreational facility frequency distributions. PCA was used to determine the number of groups that would be suitable for K Means clustering, which was then performed. The clusters with the highest average number of recreational facilities and most desirable facilities were then further evaluated for martial arts dojos using Foursquare data. The remaining neighbourhoods are shown in the Results section and their key characteristics are described.

3.2 Exploratory Analysis of the Neighbourhoods

Some neighbourhoods are large enough to be a town (i.e. Dundas), while others are collections of neighbourhoods, which will be both considered 'neighbourhoods for this analysis'. Each neighbourhood is represented by a single postal code.

Figure 1. Neighbourhood geolocations. Each blue dot is one neighbourhood.



Q: How many neighbourhoods per city are represented?

A:		
	Mississauga	*16*
	West Hamilton	*8*
	Burlington	*6*
	Oakville	*5*

3.3 Foursquare Venues

Foursquare API can be used to search and find venues available in each neighbourhood. Are daycare and childcare represented in top venues for the neighbourhoods of interest?

Figure 2. Location of top 100 Foursquare venues per neighbourhood. Each yellow dot is one venue.

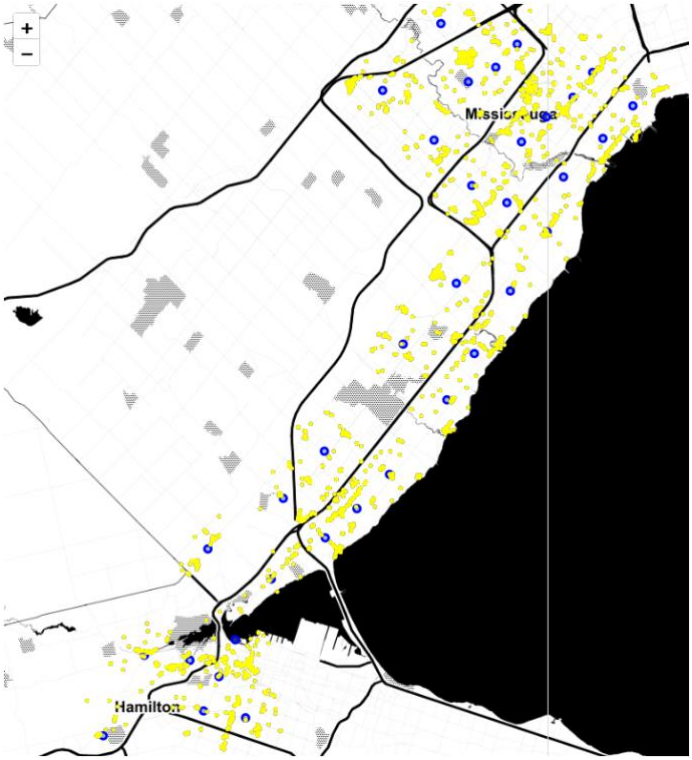
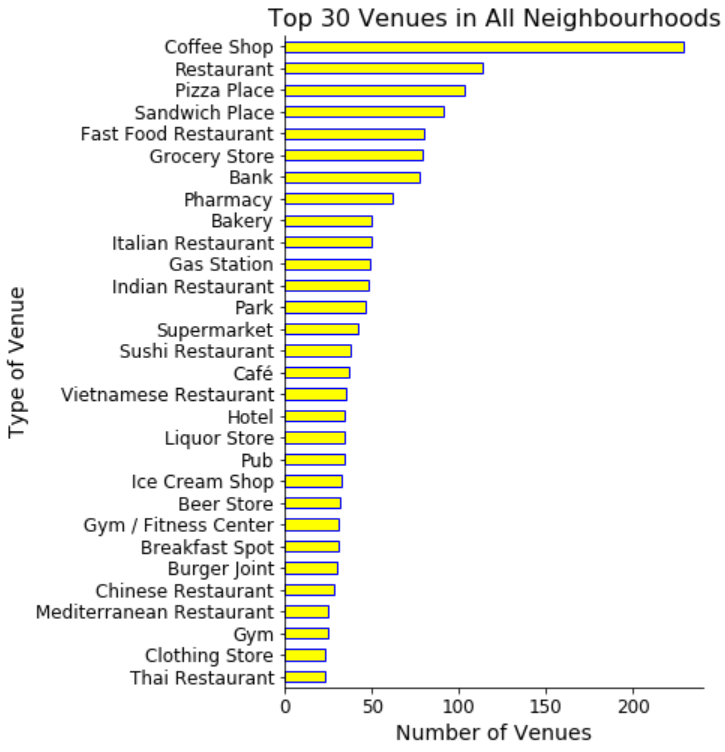


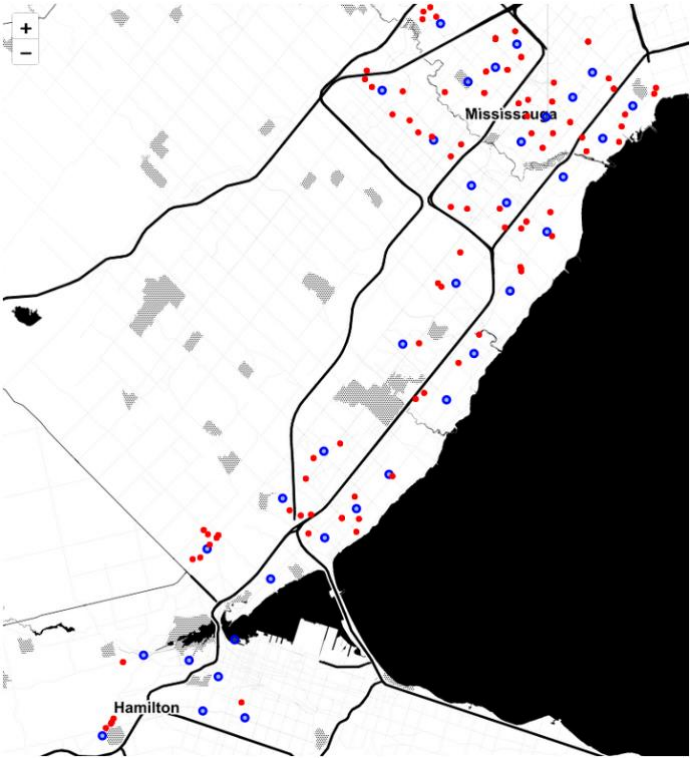
Figure 3.



3.4 Exploratory Analysis continued: Investigate Daycare and Childcare

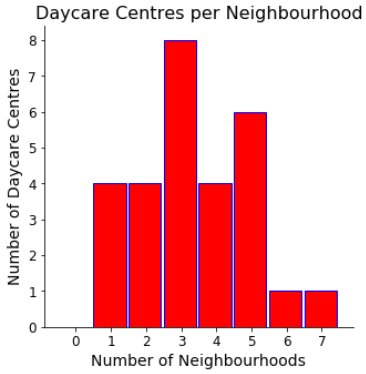
Based on the initial exploratory analysis, it is evident that Foursquare data frequency is highly influenced by restaurant data, not by daycare or schools. Due to lower frequencies of check-ins, this will have to be directly searched for. The category ID's for various establishments are available from Foursquare (URL).

Figure 3. Location of daycare centres. Each red dot is one daycare centre.



Original number of neighbourhoods in list: 35
Number of neighbourhoods that have a daycare centre: 29
Number of neighbourhoods lacking daycare which were removed: 6

Figure 4.

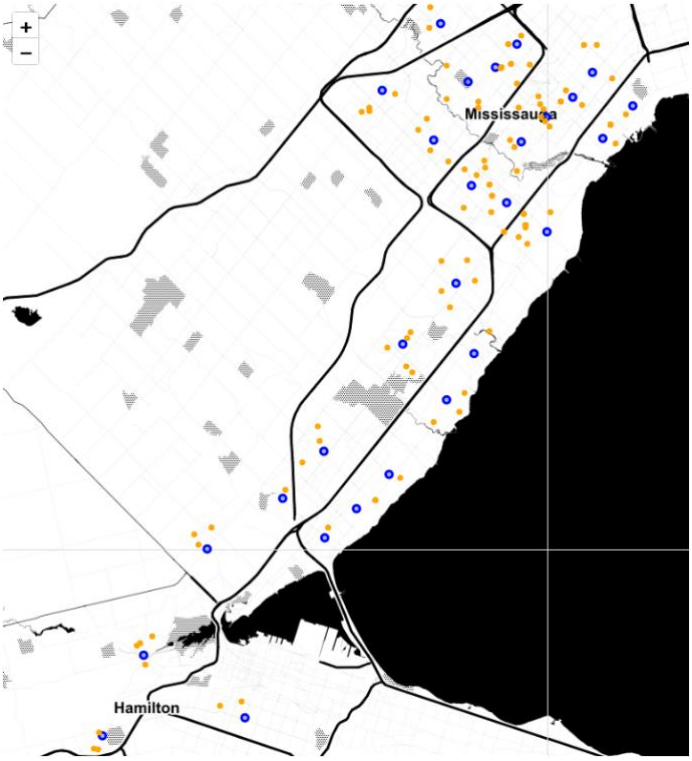


29 of the original 35 neighbourhoods remain, as they have at least one local daycare centre.

3.5 Find neighbourhoods that have a local elementary school

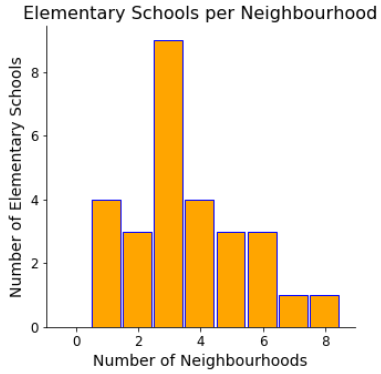
Local determined by 1 km walking distance from neighbourhood centre. Foursquare category code (URL) for elementary schools is '4f4533804b9074f6e4fb0105'.

Figure 5. Location of elementary schools. Each orange dot is one elementary school.



Original number of neighbourhoods in list:	35
Number of neighbourhoods that have a daycare and elementary school:	28
Number of neighbourhoods lacking daycare which were removed:	7

Figure 6.



28 of the original 35 neighbourhoods remain, as they have at least one local daycare centre and at least one elementary school.

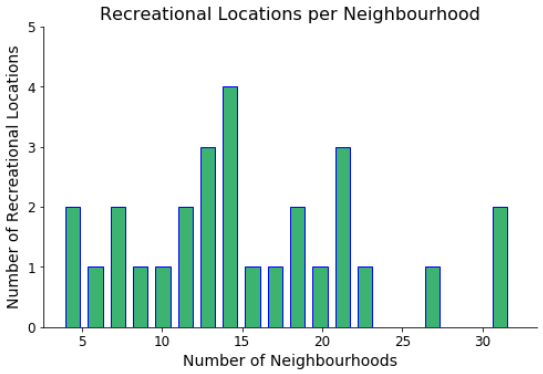
3.6 Evaluate Neighbourhood recreational quantity and diversity

Now, it is time to research the diversity of the outdoor facilities and athletic facilities available per neighbourhood. The outdoor facilities of interest to the couple are: Playground, Bike Trail, Beach, Botanical Garden, Dog Run, Fishing Spot, Lake, Nature Preserve, Park, and Pool. Furthermore, the couples are interested in athletic facilities such as: Gym / Fitness Center, Boxing Gym, Climbing Gym, Cycle Studio, Gymnastics Gym, Martial Arts Dojo, Outdoor Gym, Pilates Studio, Track, Yoga Studio, Indoor Play Area, and Recreation Center. Therefore, I will use the 'Outdoors & Recreation' category ID from Foursquare (URL).

Q: How many total unique venues categories were found in the Recreation & Outdoors search?
A: There are 38 unique recreation categories.

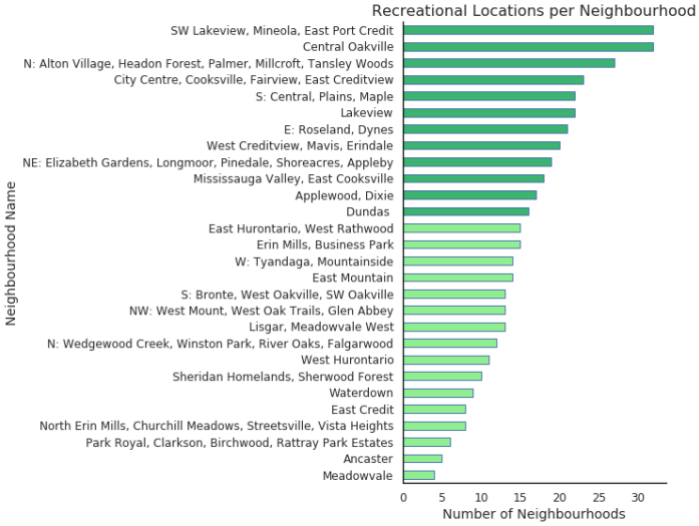
Q: How many recreational venues were found per neighbourhood?
A: There is a significant difference in recreational venues per neighbourhood, please refer to Figure 7.

Figure 7.



Q: Which neighbourhoods lead with the most recreational venues?
A: Recreational venues range from 32 to 4 per neighbourhood, please refer to Figure 8.

Figure 8.



Q: What are the most commonly found recreational venues?
A: Parks and Gyms are the 1st and 2nd most common venue, please refer to Figure 9.

Figure 9.

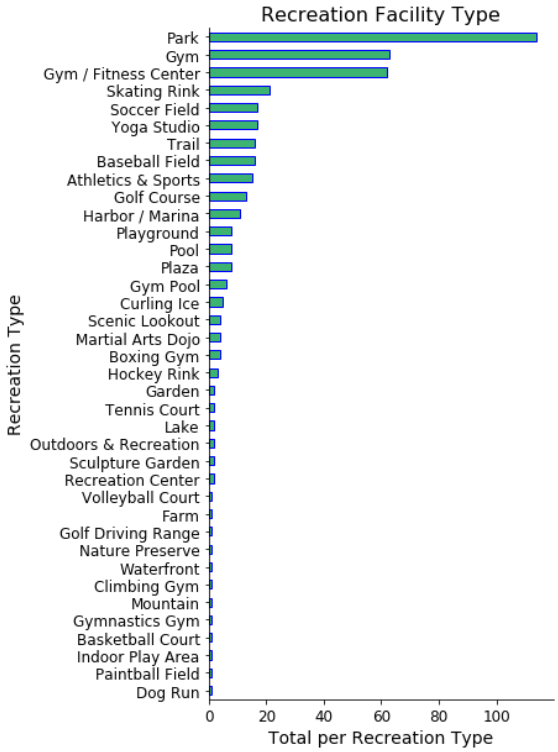
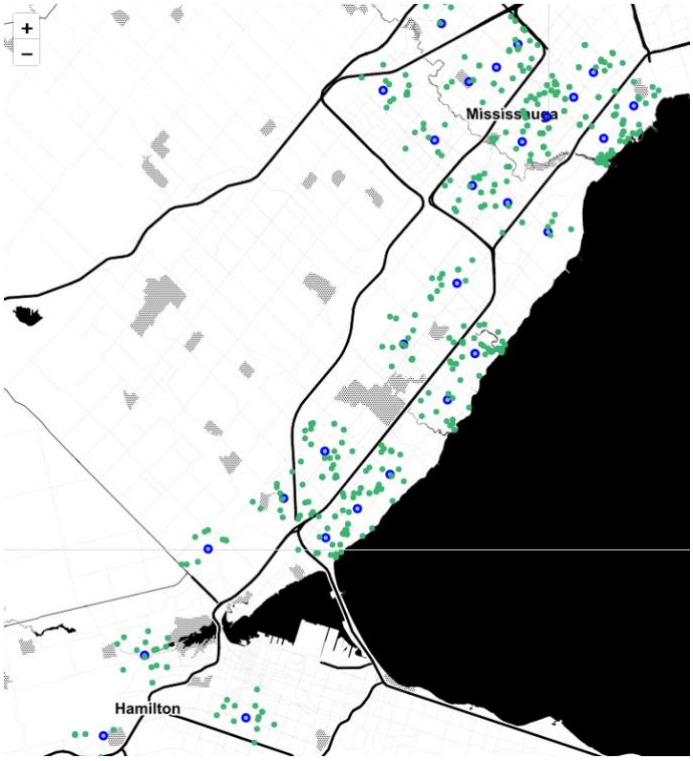


Figure 10. Location of recreational and outdoor venues. Each green dot is one venue. These neighbourhoods have already been screened to also contain at least one each of: daycare and elementary schools.



3.7 Prescriptive Analytics

Principal component analysis (PCA) is dimensionality reduction technique, which increases the interpretation of data while minimizing information loss. Therefore, we will use PCA to determine neighbourhoods that offer similar recreation facilities. First, one hot encoding allows categorical data to be manipulated as numerical data, *please refer to Figure 11*.

Figure 11. One hot encoding permits numerical analysis of categorical data. First 5 of 28 neighbourhoods shown.

	Neighbourhood	Athletics & Sports	Baseball Field	Basketball Court	Boxing Gym	Climbing Gym	Curling Ice	Dog Run	Farm	Garden	...
0	Ancaster	0.000000	0.2000	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.0	...
1	Applewood, Dixie	0.000000	0.0000	0.0	0.000000	0.058824	0.058824	0.0	0.0	0.0	...
2	Central Oakville	0.062500	0.0000	0.0	0.000000	0.000000	0.031250	0.0	0.0	0.0	...
3	City Centre, Cooksville, Fairview, East Credit...	0.043478	0.0000	0.0	0.043478	0.000000	0.000000	0.0	0.0	0.0	...
4	Dundas	0.000000	0.0625	0.0	0.000000	0.000000	0.062500	0.0	0.0	0.0	...

PCA analysis showed that 4 clusters were likely sufficient to cluster the data (**Figure 12**), this is important since K Means clustering requires the user to define the number of clusters. Indeed, 4 clusters reasonably separated the data using K Means clustering (**Figure 13**). However, there does not appear to be a regional pattern of association for these clusters (**Figure 14**).

Figure 12.

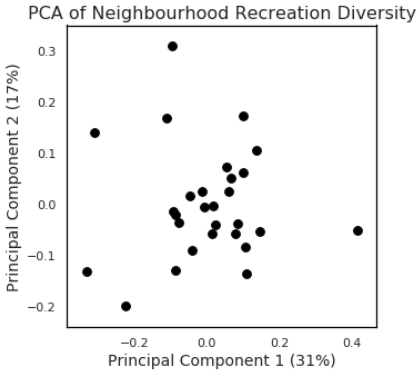


Figure 13. The larger circle is the cluster centroid.

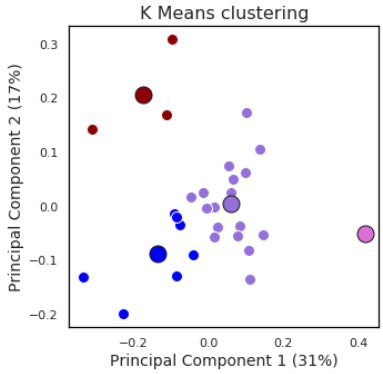
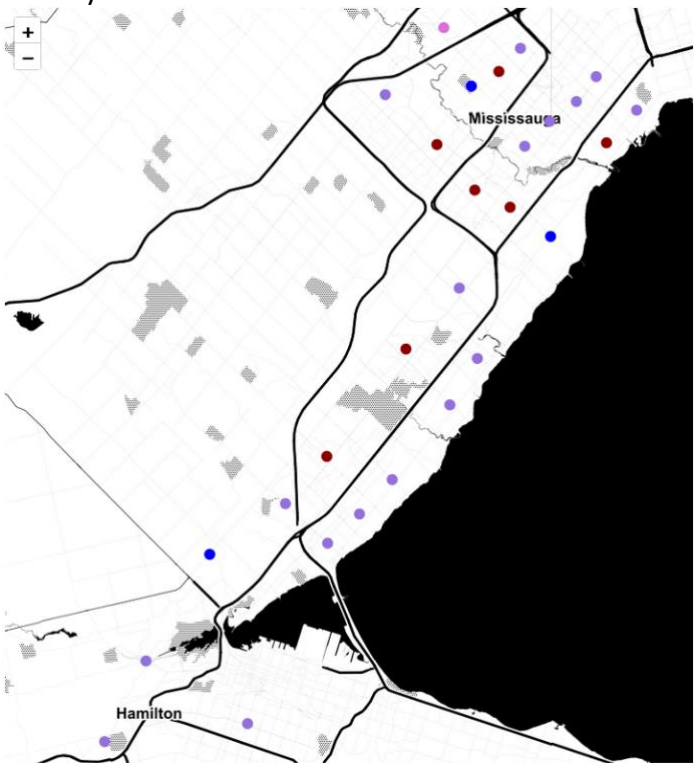


Figure 14. Location of recreational and outdoor venues. Each green dot is one venue. These neighbourhoods have already been screened to also contain at least one each of: daycare and elementary schools.



Legend:
Cluster 0 = purple
Cluster 1 = **blue**
Cluster 2 = **maroon**
Cluster 3 = pink

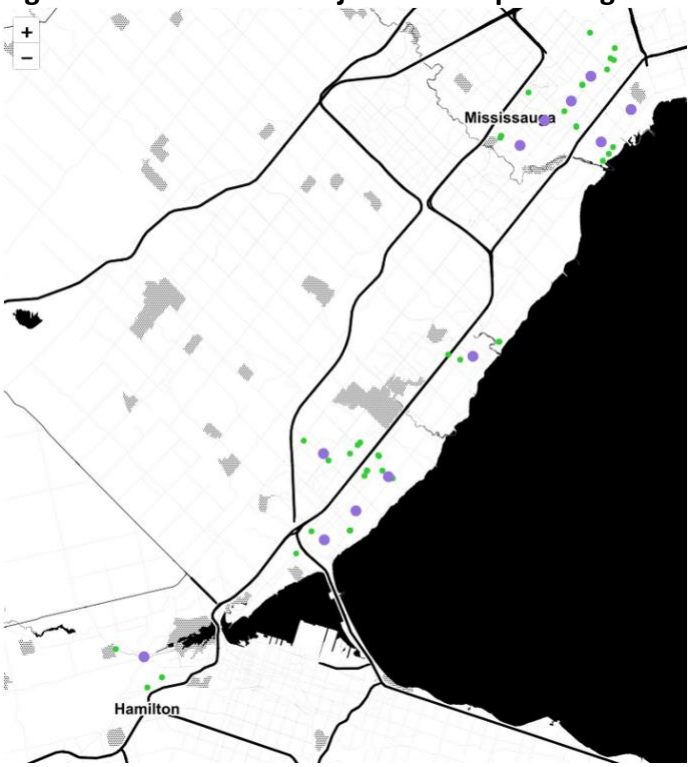
3.8 Key Features of the 4 Neighbourhood Clusters

Let's investigate the leading characteristics of the 4 different clusters, as well as the leading neighbourhoods in each cluster in terms of recreational facility quantity. In the largest cluster, **Cluster 0**, park and gyms are equally represented as the most common recreational venue, and there is a high degree of heterogeneity in the less frequent venues. Overall, this cluster has the highest quantity of recreational venues (mean = 17). The second largest cluster, **Cluster 1**, has less emphasis on parks, and a greater emphasis on gyms, golfing, yoga studios. Overall, has less than half the total venues available compared to cluster 0 and 1. However, there are more daycare centres available per neighbourhood than all other clusters. This cluster has below average number of recreational venues, the average is only 8 (range 6-9). **Cluster 2** has a high emphasis on parks, then gyms, and heterogeneous collection of individual sport venues such as: baseball, playground, boxing, martial arts, and lastly, golf. Trails, waterfront, and harbour access are also commonly emphasized. Overall, this cluster also has the highest quantity of recreational venues (mean = 17). **Cluster 3** contains only 1 neighbourhood, and prioritizes golf, gym, and gymnastics. It has the lowest total number of recreational venues: 4.

3.9 Narrowing Down Neighbourhood Clusters 0 and 2

Since there are still many suitable neighbourhoods, which meet the couple's list of requirements. The couples have decided they would also prefer a neighbourhood with a local martial arts dojo. Therefore, they wish to narrow down the neighbourhoods featured in Cluster 0 and 2, since they have to most recreational facilities. The bottom half of neighbourhoods in the remaining 24 were removed as they have fewer than 16 recreational facilities (16 is mid-way from 32 to zero). Now, 12 neighbourhoods remain. There are 43 Martial Arts Dojos in the top 12 neighbourhoods, where Lakeview was the only neighbourhood to not have a martial arts dojo, and therefore it was removed (**Figure 15**).

Figure 15. Martial arts dojos in the top 11 neighbourhoods. Each lime green dot is one dojo.



This means 11 neighbourhoods remain that satisfy all of the young couple’s criteria.

4. Results

From the analysis presented in the Methodology section, we were able to gain a significant understanding of the daycare centres, elementary schools, and recreational facilities available in each neighbourhood in the southwest region of Lake Ontario. A highlight of the top 11 desirable neighbourhoods is shown below and summarized at the end. The top 11 neighbourhoods, which meet all the criteria of the young couple, are then placed into 3 regionalized groups to improve the couple's searching for the most desirable community within these neighbourhoods.

Figure 16.

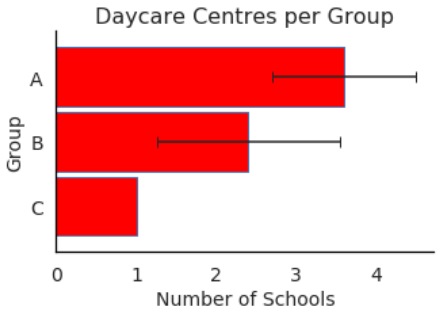


Figure 17.

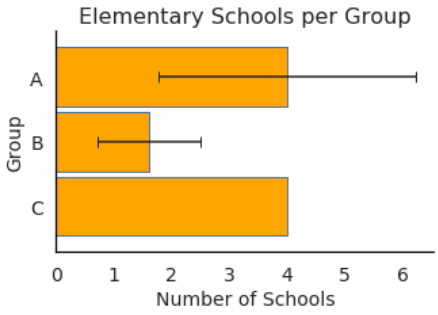
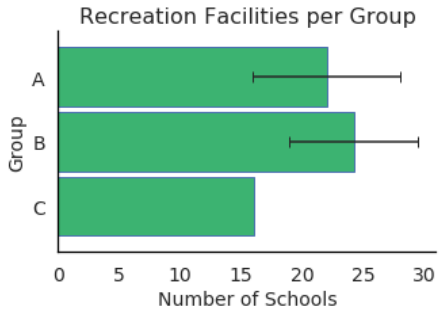
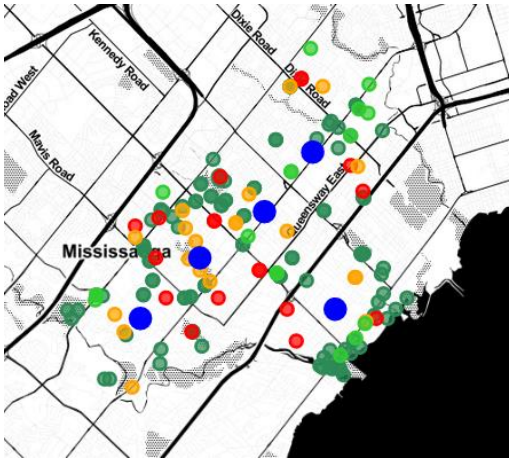


Figure 18.



Summary of Results: Three Neighbourhood Groups

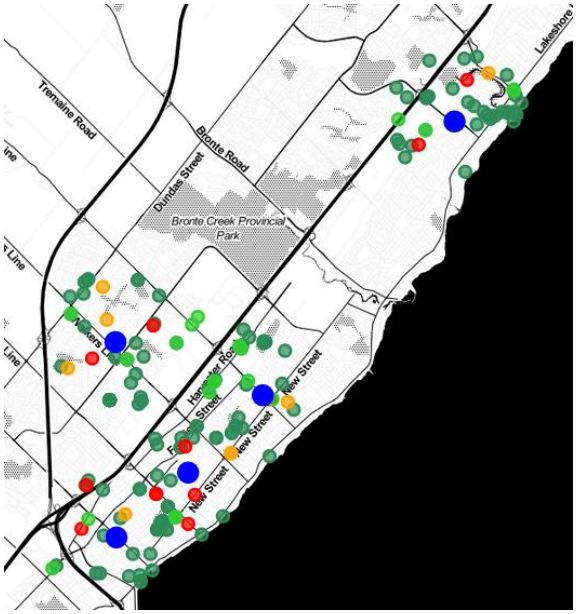
Group A: Mississauga-on-the-lake, east of the Credit River



Legend:
Neighbourhood centre = **blue**
daycare centre = **red**
elementary school = **orange**
recreational facility = **dark green**
martial arts dojo = **lime green**

The postal codes for this region are: L4Y, L5A, L5B, L5C, L5G. The neighbourhoods in this region are: Applewood, Dixie Creditview, Mississauga Valley, East Cooksville, Fairview, City Centre, Mavis, Erindale, Southwest Lakeview, Mineola, East Port Credit. This group has the highest average number of daycare centres (**Figure 16**), and number of elementary schools (**Figure 17**). It is second highest for the average number of recreational activities (**Figure 18**).

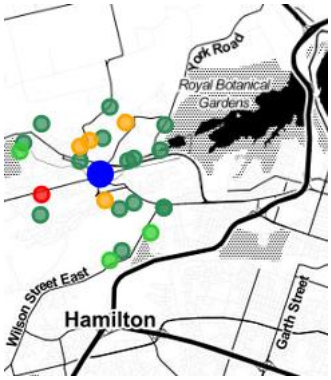
Group B: Burlington-on-the-lake, Oakville centre



Legend:
Neighbourhood centre = **blue**
daycare centre = **red**
elementary school = **orange**
recreational facility = **dark green**
martial arts dojo = **lime green**

The postal codes for this region are: L6K, L7L, L7M, L7N, L7R, L7S. The neighbourhoods in this region are: Central Oakville, Elizabeth Gardens, Longmoor, Pinedale, Shoreacres, Alton Village, Headon Forest, Palmer, Tansley Woods, Roseland, Dynes, Central, Plains, Maple. This group has the second highest average number of daycare centres (**Figure 16**), and the lowest number of elementary schools (**Figure 17**). However, it has the highest average number of recreational activities (**Figure 18**).

Group C: Dundas (west Hamilton)



Legend:
Neighbourhood centre = **blue**
daycare centre = **red**
elementary school = **orange**
recreational facility = **dark green**
martial arts dojo = **lime green**

The postal codes for this region is L9H, and the neighbourhood of this region is Dundas. This group has the lowest number of daycare centres (**Figure 16**), and the highest number of elementary schools (**Figure 17**). However, it has a quarter less recreational activities compared to the other two groups (**Figure 18**).

---

5. Discussion

Starting from 35 distinct macro-neighbourhoods, there remain 11 that meet the criteria of a young couple looking to move to the southwest shore of Lake Ontario. These 11 neighbourhoods are also clustered together and form 3 groups: Mississauga, near the lake (5 postal code regions); Burlington, near the lake, and one neighbourhood of Oakville (6 postal code regions); and Dundas city, west Hamilton (one postal code region).

These three different clusters vary slightly in the abundance of the key elements that the couple are looking for: (1) local daycare, (2) local elementary school, (3) high quantity and quality of recreational activities, and (4) an additional item; a local Martial Arts Dojo. Since 11 neighbourhoods meet the criteria, the couple can now focus on the particular recreational activities they prefer in the smaller communities within each of these neighbourhoods. This should save time and resources in approaching neighbourhood research as well as time spent at open houses and community events of neighbourhoods that are overall not a good fit for this particular couple.

Having three different large clusters of neighbourhoods is helpful, as it affords the couple more flexibility in finding a neighbourhood which also optimizes their working commute. This is largely the reason that the couple chose to not yet further reduce the number of desirable neighbourhoods.

Other factors may play a role in the search of the ideal neighbourhood, such as housing costs, which was not factored into the present analysis. Housing costs can vary within communities due to the proximity to desirable features, waterfront and parks, or undesirable locations such as waste management facilities and high-traffic roads. Therefore, while this analysis provides a more focussed approach for starting to describe ideal neighbourhoods, further investigation is warranted to understand the regional differences within communities of each desirable neighbourhood.

This analysis was conducted using data analytics provided by the Foursquare API. It is possible that daycare centres that are small or do not have an online presence could be missing from this analysis, which would exclude an entire neighbourhood. Foursquare data is also user generated, and so it is possible that the presence and absence of daycares centres and other similar facilities is different today during a global pandemic where multiple businesses are still currently closed. The couple will run this data analysis again later when more businesses open up, to see if the results change.

---

6. Conclusion

In this present study, we used a data science approach to explore and select neighbourhoods based on the preferences of a young couple looking to buy their first home. We collected existing data using a postal code database as well as the Foursquare API. Based on these data and the ensuing data analytics, we were able to provide valuable information which would assist the couple in narrowing their search for a desirable neighbourhood.

We recognize this study is based on limited data and is also uniquely impacted by a global pandemic, which affects user generated data. However, this study provides a good initial starting point for considering the qualities of a desirable neighbourhood for a young couple who appreciate local amenities and outdoor activities. This analysis can be adapted to other cities around the globe by adapting the code presented above, and can be used for all sorts of factors that people require when considering a neighbourhood that they would like to live in.