Identifying Neighbourhoods in Toronto based on Childcare Support

J
December 2020
Project: Applied Data Science Capstone

1. Introduction

Young families who have newborn children or are expecting a child often decide to move into a larger living space in a family friendly neighbourhood to accommodate their growing family. This often involves buying or moving into a new home in a neighbourhood that has services and facilities to support. One service that is especially important for young families to consider when moving to a new neighbourhood is access to childcare services.

Childcare services provide parents an option to return to work and provide infants, toddlers, and young children an avenue to learn, socialize, and play. Having access to these services is critical for families to be able to pursue their careers and provide for their family, while ensuring their young children can develop appropriately.

Understanding the availability and capacity of childcare services can be an important factor in determining what neighbourhood to live in. For Toronto, childcare service capacity is limited and varies considerably by neighbourhood. Reserving space in childcare centres of choice often requires parents to enroll their child when they know they have conceived. This means parents with young children or who are in the process of family planning should take childcare service accessibility and quality into consideration before moving to a new neighbourhood, even if a child is still in the future.

The goal of this project is to identify Toronto neighbourhoods based on the capacity of childcare services available. This will support in decision making for young families planning to move to accommodate their growing family. Additionally, it can be used to identify areas where childcare services are underrepresented.

2. Data

2.1 Sources

Neighbourhood data on the City of Toronto will be accessed using the [City of Toronto's Open Data Portal](#) using the [Neighbourhoods](#) data set. This data set includes the boundaries of each neighbourhood and will be used to identify the neighbourhood each childcare facility belongs to. The data will be imported both as a GEOJSON file to be used to visualize the analysis on a choropleth map once finalized. Additionally the neighbourhood data will be extracted as a CSV to be used in data cleaning for other parts of the analysis.

Childcare data will be pulled from the City of Toronto's Open Data Portal using the [Licensed Child Care Centres](#) data report. In this data lists all licensed childcare service facilities in Toronto and includes

information on their location, geo coordinates, capacity by age group, as long with other descriptive data. This will be used as the list of childcare service providers as part of this analysis and used in grouping them by their respective neighbourhood.

Toronto neighbourhoods have varying geographical areas and shapes, with vastly different populations. This may skew identifying neighbourhoods as family friendly as neighbourhoods with larger populations may naturally have more capacity for childcare services. To normalize, we will pull the most recent population census data by neighbourhood to create a ratio of childcare service capacity by population. This will allow us to compare young family neighbourhoods without natural bias based on population differences between neighbourhoods. The latest population census data is from 2016 and can be accessed in the [Neighbourhood Profiles](#) data set.

2.2 Cleaning the data

The Toronto Neighbourhood Data primary use was the defining the boundaries of each of the 140 neighbourhoods in Toronto city limits. This did not require cleaning of the data as the GEOJSON file had the necessary coordinate data for the Choropleth map to be used later on.
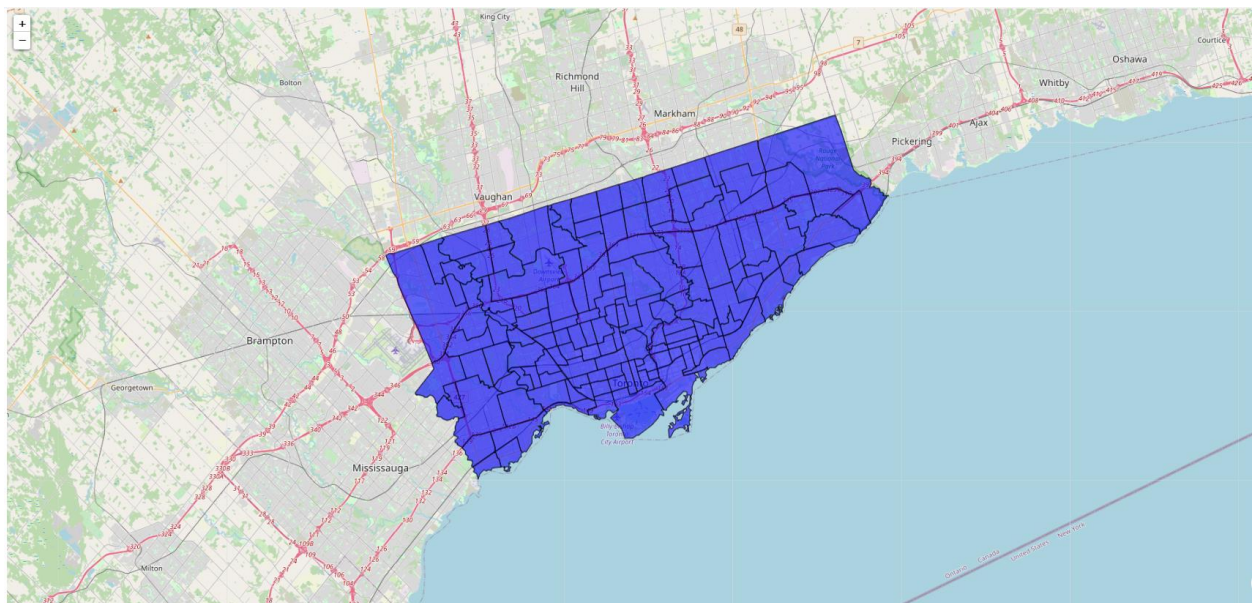
The secondary use of the Toronto Neighbourhood Data was to define the specific neighbourhoods and their names to be used in the analysis. To accomplish this, the JSON file was converted into a data frame and only the necessary columns were kept: neighbourhood ID, neighbourhood code, and neighbourhood name. This produced 140 unique neighbourhoods for the city of Toronto with a data frame that had a shape of (140, 3). A sample of the data frame can be seen below.

	neighbourhood_id	neighbourhood_code	neighbourhood_name
0	9801	96	Casa Loma
1	9802	95	Annex
2	9803	109	Caledonia-Fairbank
3	9804	64	Woodbine Corridor
4	9805	103	Lawrence Park South
5	9806	130	Milliken
6	9807	53	Henry Farm
7	9808	26	Downsview-Roding-CFB
8	9809	6	Kingsview Village-The Westway
9	9810	124	Kennedy Park

Next, importing and cleaning the Licensed Child Care Centres data was needed to extract the information that will be used in our clustering analysis later on to determine which neighbourhoods provide more childcare services for young families. The Licensed Child Care Centres data has 1037 service providers listed with 22 features (columns) of information for each one. This information included the geo coordinates of each provider. However, it did not include the neighbourhood in which the childcare service provider was located in.

To add the neighbourhood, a function was created using the **Shapely** python library to pass in the geo coordinates of each provider and identify what neighbourhood they were in utilizing the GEOJSON file of Toronto Neighbourhoods. This information was then added to the Childcare Services data frame, resulting in 23 features.

As most of the features would not be needed for the analysis, all unnecessary features were removed resulting in a data frame with the shape (1037, 9). The features that remained were:

- **provider_id**: unique provider ID
- **provider_name**: provider name
- **neighbourhood_name**: neighbourhood name
- **infant_cap**: capacity for children 0-18 months
- **toddler_cap**: capacity for children 18-30 months
- **preschooler_cap**: capacity for children +30 months to 1st grade
- **kindergarten_cap**: capacity for children in full-day kindergarten
- **gradeschooler_cap**: capacity for children in 1st grade plus
- **total_cap**: capacity for all childcare services

Having a final list of information that will be used in our clustering analysis later on we can run light analysis of the overall data set. Looking at the descriptive statistics of the childcare service, we can see that mean total capacity for all age groups of the typical service provider is 75.9 spaces available, with a standard deviation of 48.8. Additionally, the capacity for infant and toddler care with a mean of 3.8 and 10.6 respectively, is much lower than that of preschool age and above. This means that families looking for infant and toddler care must be cognizant of the limited space available for these services.

The shape of the data frame is: (1037, 9)

	infant_cap	toddler_cap	preschooler_cap	kindergarten_cap	gradeschooler_cap	total_cap
count	1037.000000	1037.000000	1037.000000	1037.000000	1037.000000	1037.000000
mean	3.779171	10.569913	23.418515	14.555448	23.614272	75.937319
std	6.086128	10.818982	18.390140	20.446230	31.483705	48.779990
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	16.000000	0.000000	0.000000	41.000000
50%	0.000000	10.000000	24.000000	0.000000	0.000000	62.000000
75%	10.000000	15.000000	32.000000	26.000000	40.000000	98.000000
max	30.000000	86.000000	180.000000	117.000000	225.000000	342.000000

3. Methodology

3.1 Exploratory Data Analysis

Looking at the Total Capacity for each neighbourhood we can see there is a large deviation in the available childcare services from the top 10 neighbourhoods to the bottom 10 neighbourhoods.

**Top 10 Toronto Neighbourhoods by Total Capacity of Childcare Services: All Age Groups**

neighbourhood_name	infant_cap	toddler_cap	preschooler_cap	kindergarten_cap	gradeschooler_cap	total_cap
Islington-City Centre West	110	270	546	194	444	1564
Waterfront Communities-The Island	157	295	599	135	245	1431
Parkwoods-Donalda	30	115	304	247	500	1196
Mimico	15	127	235	375	423	1175
Trinity-Bellwoods	46	120	320	273	393	1152
High Park-Swansea	40	135	370	221	375	1141
Banbury-Don Mills	46	153	392	143	339	1073
Downsview-Roding-CFB	70	129	327	232	307	1065
Kensington-Chinatown	100	205	397	178	175	1055
East End-Danforth	26	195	254	216	363	1054

Bottom 10 Toronto Neighbourhoods by Total Capacity of Childcare Services: All Age Groups

neighbourhood_name	infant_cap	toddler_cap	preschooler_cap	kindergarten_cap	gradeschooler_cap	total_cap
Briar Hill-Belgravia	0	15	24	0	0	39
Pelmo Park-Humberlea	10	10	16	13	15	64
Rustic	16	20	32	0	0	68
Beechborough-Greenbrook	20	45	52	0	0	117
Caledonia-Fairbank	10	15	16	26	60	127
Morningside	10	25	62	12	28	137
Centennial Scarborough	0	14	20	78	30	142
Newtonbrook East	30	40	39	16	21	146
Etobicoke West Mall	20	25	56	26	30	157
Thorncliffe Park	10	15	56	0	81	162

For easy visualization of the total childcare service capacity neighbourhood, we can add this information to a Choropleth map. This does not yet take into consideration population per neighbourhood, which may skew the initial output to those with larger populations.

Toronto Neighbourhoods Total Capacity of Childcare Services: All Age Groups
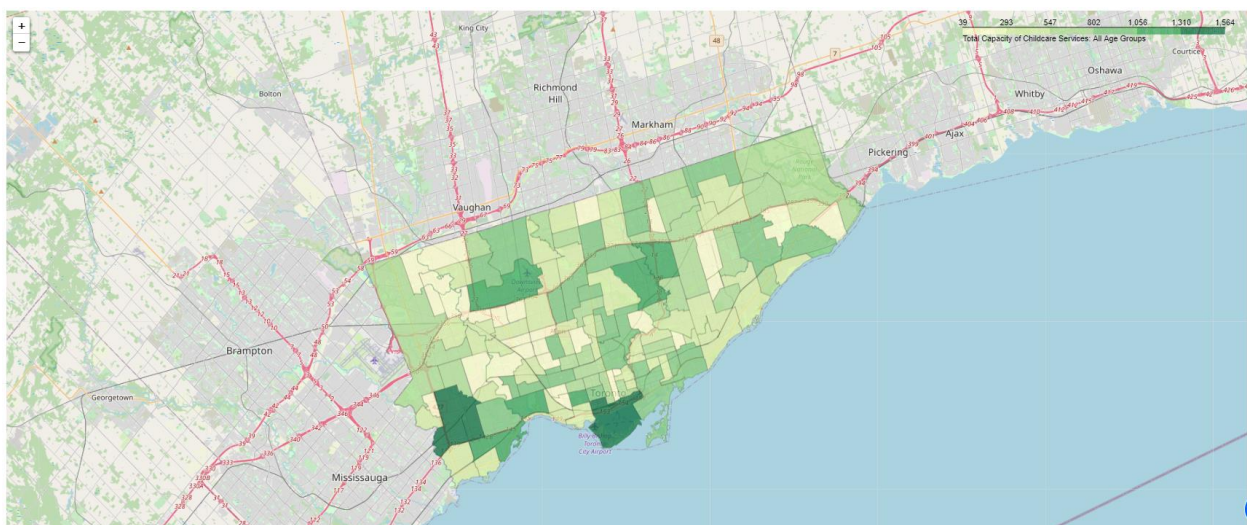

To normalize the capacity by population, we will need to layer in Toronto population census data using the Neighbourhood Profiles data set from Open Data Toronto. This data includes census information by neighbourhood large number of "Characteristics". As we are only interested in the 2016 Population data, we will only keep this information and remove all the rest. The 2016 population data by neighbourhood will then be appended to the Toronto Childcare Service information we have already cleared.

Looking at the ratio of Total Capacity of Childcare Services to Population, the Choropleth map changes considerably. It shows different groups of neighbourhoods that have higher ratios of Childcare services, which could help identify neighbourhoods for Young Families.

Toronto Neighbourhoods Total Capacity of Childcare Services to Population Ratio: All Age Groups
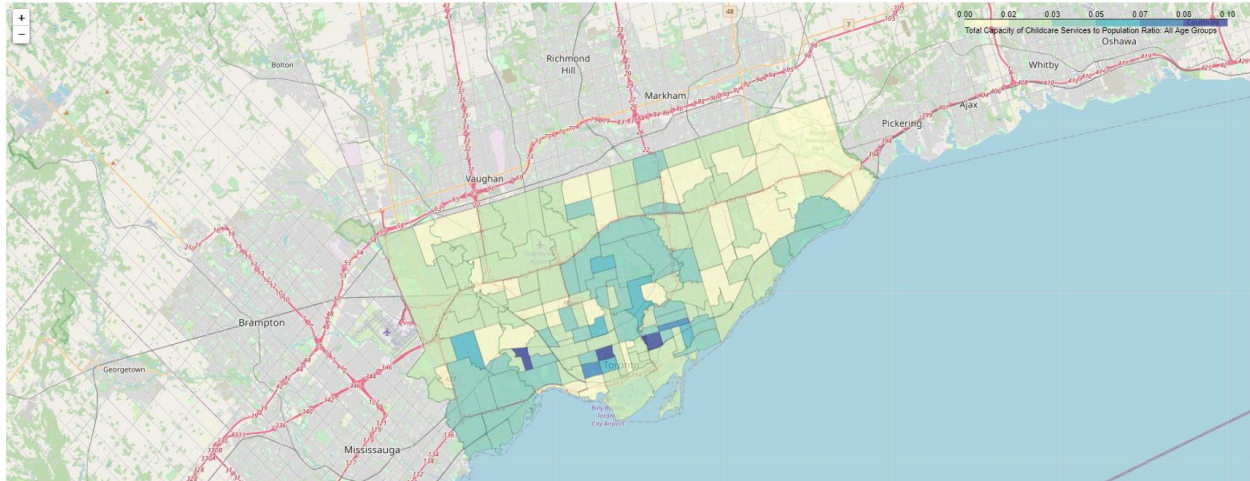
Comparing the neighbourhoods can also be visualized using a scatter plot to see similarities across neighbourhoods by their capacity to population ratios. The 5-dimensional chart below, plots the different age group capacity to population ratios with three youngest age groups represented by the x, y, and z axis and the two older one by size and color. Neighbourhoods with lower childcare capacity to population ratios are represented by small purple dots close to the bottom-left corner. Neighbourhoods with higher capacity to population ratios appear are larger yellow/green dots that are closer to the top-right corner.
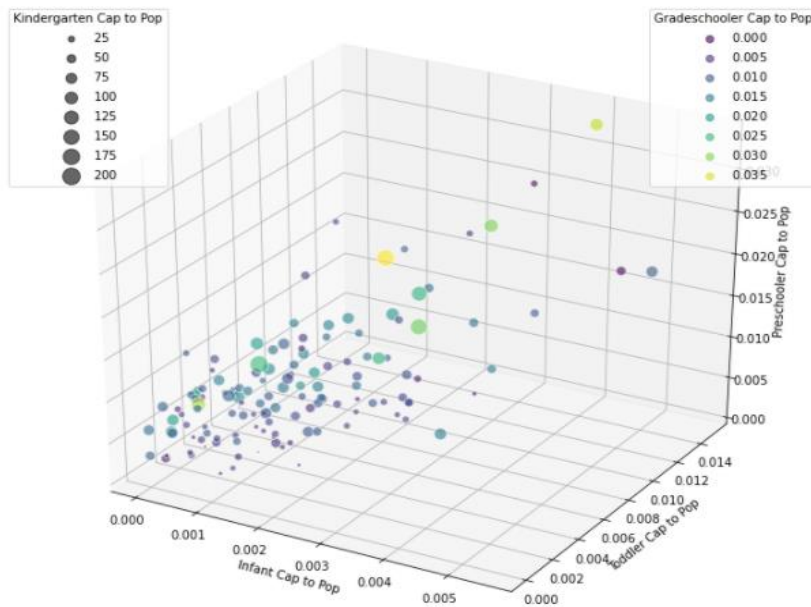
3.2 K-means Clustering

The purpose of this analysis is to help identify neighbourhoods based on the childcare service capacity they offer for young families. K-means Clustering identifies data points based on the distance between the data point and its nearest neighbouring data points. This helps identifies data points with commonalities amongst them and works well with our analysis as we are only utilizing one-dimensional numeric data.

The result of the clustering applied to the childcare service data will be neighbourhood clusters based on childcare service capacity for the varying age groups. For the purposes of this analysis we will use 5 clusters to label the neighbourhoods. Once we have cluster labels for each of the neighbourhoods, we will be able to identify which ones offer higher level of childcare service capacity to compare to others. This can ultimately be used to help identify which neighbourhoods have more capacity to support young families.

Utilizing the k-means model in the scikit-learn python library with the neighbourhood data, 5 distinct clusters were identified. Examining the descriptive statistics of each cluster, we can apply a "Cluster Name" to each of the clusters to describe the general capacity of childcare services for neighbourhoods in the cluster. This was done by primarily looking at the **mean capacity to population** ratios across each of the age groups in the cluster. From this each cluster was given a name ordered from the lowest total capacity to population mean to highest.

Cluster Assigned	Cluster Name	Total Cap to Pop Ratio - Mean
1	Low Childcare Support	0.0157
2	Low to Moderate Childcare Support	0.0287
3	Moderate and Balanced Childcare Support	0.0418
4	High Newborn Childcare Support	0.0452
5	High Childcare Support	0.0765

*See Appendix A for complete details

4. Results

Adding the labels to each of the neighbourhoods, a Choropleth map of Toronto neighbourhoods can now be created with each of the Clusters Assigned. In the Choropleth map each color represents a cluster with yellow representing cluster 1 and dark green representing cluster 5.

Toronto Neighbourhoods by Childcare Service K-means Clusters
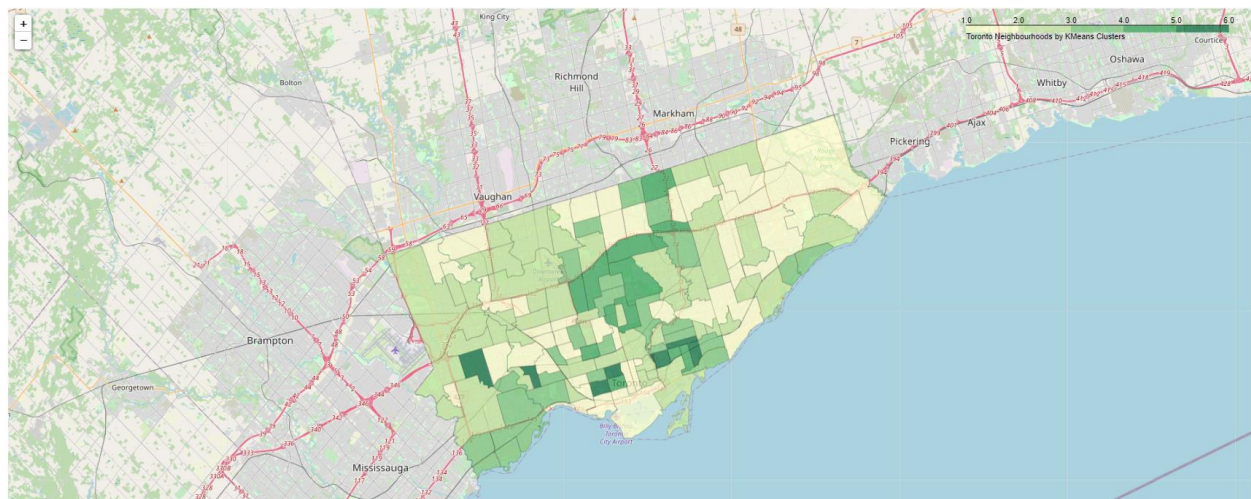
Examining the Choropleth map above, it is apparent that there are neighbourhoods with higher support for childcare services than others.

In the immediate downtown neighbourhoods of Church-Yonge Corridor, Baystreet Corridor, Waterfront Communities and Islands, there is low support for childcare services. Once you move away from the downtown core neighbourhoods the support for childcare services increases. Specifically, pockets of increased childcare service capacity by population exist in neighbourhoods:

- Directly east of the city and Don Valley Parkway
- West of the city close to High Park, Mimico, and the Junction, and
- North / North East of the city

As you move Northwest and to the neighbourhoods on the boundaries of city limits, childcare support services show lower support. Many of the neighbourhoods in these areas fall under the clusters "Low Childcare Support" or "Low to Moderate Childcare Support".

Looking at the 5-dimensional plot again and substituting the "Cluster Labels" for the "Grade-schooler to Cap Ratio" we can see the clusters largely align with the initial exploratory analysis.

Smaller dots appear more often in the bottom left corner of the plot (neighbourhoods with lower cap to pop ratios) while the larger dots appear closer to the top right corner.
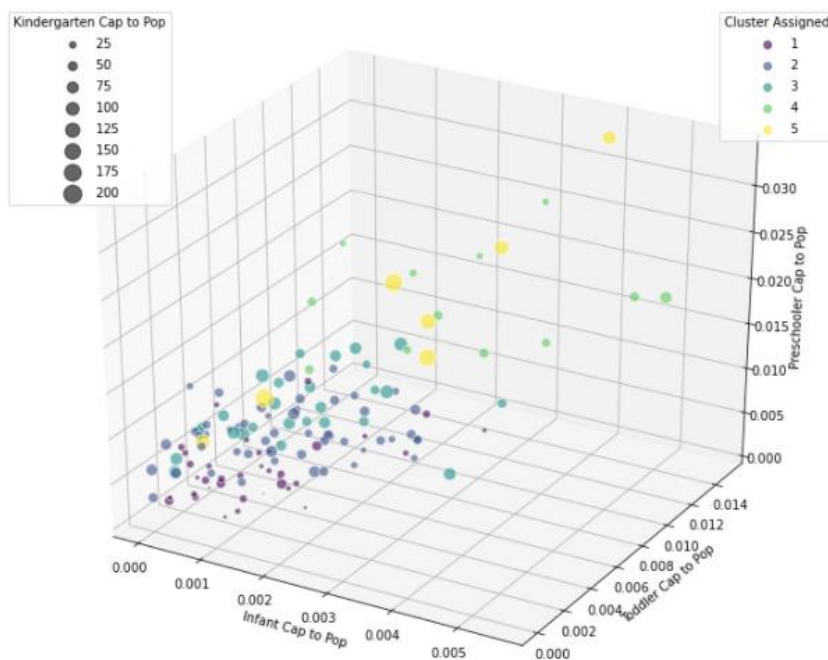
Using the k-means clustering, we have successfully able to identify neighbourhoods in Toronto based on their childcare service capacity. This can help in the decision support for Young families as they identify which areas in Toronto to move to.

5. Discussion

For Young Families planning on moving and need childcare services, this analysis can be used to identify the neighbourhoods that have the most capacity to support them. Neighbourhoods that appear in clusters 3-5 will likely have the childcare services needed to support a young family at varying age

groups. Even neighbourhoods in cluster 3 are generally bordering neighbourhoods that have higher capacity. This information can support decisions on where to live for those that need to consider childcare services carefully in their planning.

Additionally, this information could be used to help identify neighbourhoods that are potentially underserviced for childcare services. Those interested in setting up a new Licensed childcare service to support underserviced neighbourhoods in Toronto could layer this information on top of other demographic, socioeconomic, and other data to help identify the best neighbourhoods for that support.

Conclusion

Deciding on what neighbourhood to move to or potentially buy a home in is not always a straightforward decision. It requires examining a variety of factors that align with individual needs and lifestyle of those choosing to make move. Young families in Toronto who are often faced with the decision to move to meet their growing family, want to be informed on areas that offer the childcare service support that they need. The analysis here can help them narrow their search by identifying neighbourhoods that have the highest capacity of childcare services and support their decision on where to live in Toronto.

Disclaimers

This project is limited in scope to only include publicly accessible information made available through the City of Toronto and other digital resources.

It is a one-dimensional analysis and only investigates the capacity of listed childcare service providers to determine Identify Family Neighbourhoods. It does not take into consideration other factors that impact the access to childcare services in a neighbourhood including childcare service providers ratings, income, education, socioeconomics, demographics, other public facilities and many other important social determinants.

Appendix A: Jupyter Notebook Descriptive Statistics of each Cluster

Cluster: "Low Childcare Support"

This cluster has the lowest capacity of childcare services by population across all capacity to ratio across all neighbourhoods in this cluster. Neighbourhoods in this cluster may not have the capacity needed for young families or are underserviced for the neighbourhoods childcare needs.

In [32]: ⊙ df_tor_ccs_nbh_merge.loc[df_tor_ccs_nbh_merge['Cluster Assigned'] == 1].describe()

Out[32]:

	infant_cap	toddler_cap	preschooler_cap	kindergarten_cap	gradeschooler_cap	total_cap	pop_2016	inf_ctp	tod_ctp	pre_ctp	kin_ctp	gra_ctp	tot_ctp	Cluster	Cluster Assigned
count	45.000000	45.000000	45.000000	45.000000	45.000000	45.000000	45.000000	45.000000	45.000000	45.000000	45.000000	45.000000	45.000000	45.0	45.0
mean	24.644444	59.755556	125.800000	58.622222	91.000000	359.822222	22609.977778	0.001141	0.002680	0.005493	0.002571	0.003767	0.015653	1.0	1.0
std	27.957887	51.719592	107.836282	47.561582	77.085254	265.596694	13175.382933	0.000889	0.001535	0.002826	0.001784	0.001979	0.005588	0.0	0.0
min	0.000000	10.000000	16.000000	0.000000	0.000000	39.000000	6577.000000	0.000000	0.000711	0.001492	0.000000	0.000000	0.002735	1.0	1.0
25%	10.000000	25.000000	56.000000	26.000000	30.000000	178.000000	13362.000000	0.000450	0.001527	0.003237	0.001431	0.002395	0.012169	1.0	1.0
50%	20.000000	42.000000	96.000000	52.000000	75.000000	279.000000	18675.000000	0.001071	0.002413	0.004760	0.002510	0.003717	0.015718	1.0	1.0
75%	30.000000	80.000000	152.000000	91.000000	105.000000	436.000000	26274.000000	0.001609	0.002975	0.007596	0.003854	0.005245	0.019583	1.0	1.0
max	157.000000	295.000000	599.000000	206.000000	330.000000	1431.000000	65913.000000	0.003799	0.006842	0.012133	0.007217	0.007597	0.027139	1.0	1.0

Cluster: "Low to Moderate Childcare Support"

This cluster has a balanced capacity of childcare services across all age groups that grows progressively with each age group.
However, the cluster has lower **mean capacity to population** ratios compared to clusters other clusters.
Neighbourhoods have capacity at all age groups to support young families but may not be able to support as many families overall as other neighbourhoods.

```
df_tor_ccs_nbh_merge.loc[df_tor_ccs_nbh_merge['Cluster Assigned'] == 2].describe()
```

[33]:

	infant_cap	toddler_cap	preschooler_cap	kindergarten_cap	gradeschooler_cap	total_cap	pop_2016	inf_ctp	tod_ctp	pre_ctp	kin_ctp	gra_ctp	tot_ctp	Cluster	Cluster Assigned
count	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.0	49.0
mean	28.857143	77.306122	169.469388	119.265306	185.081633	579.979592	20049.551020	0.001345	0.003768	0.008260	0.005991	0.009298	0.028662	3.0	2.0
std	25.117225	48.713107	97.851312	56.927503	83.629350	276.796066	8504.254759	0.000952	0.001537	0.002530	0.001649	0.001908	0.004252	0.0	0.0
min	0.000000	0.000000	24.000000	36.000000	65.000000	166.000000	7865.000000	0.000000	0.000000	0.003051	0.002648	0.005292	0.021106	3.0	2.0
25%	10.000000	45.000000	88.000000	78.000000	116.000000	366.000000	13593.000000	0.000556	0.002955	0.006725	0.005051	0.007909	0.025981	3.0	2.0
50%	26.000000	70.000000	170.000000	111.000000	180.000000	546.000000	17757.000000	0.001243	0.003790	0.008024	0.005907	0.009201	0.028316	3.0	2.0
75%	36.000000	95.000000	216.000000	156.000000	240.000000	733.000000	27446.000000	0.001997	0.004685	0.009865	0.006619	0.010710	0.031418	3.0	2.0
max	110.000000	270.000000	546.000000	274.000000	444.000000	1564.000000	43965.000000	0.003220	0.006498	0.014154	0.009829	0.012855	0.039478	3.0	2.0

Cluster: "Moderate and Balanced Childcare Support"

This neighbourhood has a more balanced capacity of childcare services by population looking at the *tot_ctp* ratio across all neighbourhoods in this cluster.

```
df_tor_ccs_nbh_merge.loc[df_tor_ccs_nbh_merge['Cluster Assigned'] == 3].describe()
```

[34]:

	infant_cap	toddler_cap	preschooler_cap	kindergarten_cap	gradeschooler_cap	total_cap	pop_2016	inf_ctp	tod_ctp	pre_ctp	kin_ctp	gra_ctp	tot_ctp	Cluster	Cluster Assigned
count	27.000000	27.000000	27.000000	27.000000	27.000000	27.000000	27.00000	27.000000	27.000000	27.000000	27.000000	27.000000	27.000000	27.0	27.0
mean	22.703704	79.296296	177.037037	153.925926	275.888889	708.851852	17182.00000	0.001360	0.004668	0.010240	0.008987	0.016527	0.041781	2.0	3.0
std	16.797012	42.031870	91.151388	72.067237	101.063321	281.116962	7246.79855	0.001015	0.001766	0.003360	0.002105	0.003110	0.006642	0.0	0.0
min	0.000000	20.000000	62.000000	52.000000	131.000000	315.000000	7804.00000	0.000000	0.001518	0.004250	0.005157	0.012454	0.031022	2.0	3.0
25%	10.000000	45.000000	95.000000	105.500000	202.500000	456.000000	11280.50000	0.000846	0.003363	0.007660	0.006953	0.014494	0.037517	2.0	3.0
50%	20.000000	80.000000	179.000000	130.000000	247.000000	641.000000	16472.00000	0.001281	0.004874	0.009680	0.009237	0.015674	0.040908	2.0	3.0
75%	38.000000	110.000000	233.500000	208.000000	355.500000	955.000000	21474.00000	0.001828	0.005662	0.012620	0.010410	0.018664	0.044735	2.0	3.0
max	66.000000	195.000000	370.000000	375.000000	500.000000	1196.000000	34805.00000	0.003946	0.009120	0.016553	0.012360	0.024269	0.057404	2.0	3.0

Cluster: "High Newborn Childcare Support"

This cluster has the **highest childcare services mean** for infant, toddler, and preschool looking at the capacity to population ratios across all clusters.
However, this neighbourhood does not have as many services available for kindergarten and gradeschooler age groups.
Therefore it is more heavily weighed toward supporting new born children and families that are having their children from age 3 months to 4 years.

```
df_tor_ccs_nbh_merge.loc[df_tor_ccs_nbh_merge['Cluster Assigned'] == 4].describe()
```

[35]:

	infant_cap	toddler_cap	preschooler_cap	kindergarten_cap	gradeschooler_cap	total_cap	pop_2016	inf_ctp	tod_ctp	pre_ctp	kin_ctp	gra_ctp	tot_ctp	Cluster	Cluster Assigned
count	12.000000	12.000000	12.000000	12.000000	12.000000	12.000000	12.000000	12.000000	12.000000	12.000000	12.000000	12.000000	12.000000	12.0	12.0
mean	43.750000	136.416667	316.916667	90.083333	125.500000	712.666667	15697.750000	0.002982	0.008978	0.020225	0.005736	0.007294	0.045214	4.0	4.0
std	22.291968	49.074078	111.597050	40.214105	79.198829	240.116765	4641.711697	0.001447	0.002771	0.003401	0.001853	0.004460	0.006551	0.0	0.0
min	20.000000	64.000000	170.000000	35.000000	0.000000	347.000000	7727.000000	0.000861	0.004439	0.014982	0.003357	0.000000	0.032462	4.0	4.0
25%	30.000000	103.750000	237.750000	59.500000	71.250000	463.500000	13716.000000	0.002066	0.007392	0.016866	0.004229	0.004947	0.042617	4.0	4.0
50%	39.500000	134.000000	300.500000	84.500000	155.000000	774.000000	16249.000000	0.002717	0.009093	0.021885	0.005799	0.007147	0.045859	4.0	4.0
75%	49.500000	156.250000	397.000000	117.000000	179.250000	872.000000	17845.250000	0.003716	0.011090	0.022133	0.006775	0.010565	0.048355	4.0	4.0
max	100.000000	230.000000	514.000000	178.000000	225.000000	1055.000000	23236.000000	0.005573	0.013814	0.026441	0.009919	0.013413	0.058791	4.0	4.0

Cluster: "High Childcare Support"

This cluster has a high ratio of capacity of childcare services across all age groups of young children.
Neighbourhoods in this cluster have the highest total capacity by population and would have services available to support young families from new born children to 12 year olds in Before and After school programs.

```
df_tor_ccs_nbh_merge.loc[df_tor_ccs_nbh_merge['Cluster Assigned'] == 5].describe()
```

[36]:

	infant_cap	toddler_cap	preschooler_cap	kindergarten_cap	gradeschooler_cap	total_cap	pop_2016	inf_ctp	tod_ctp	pre_ctp	kin_ctp	gra_ctp	tot_ctp	Cluster	Cluster Assigned
count	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.0	7.0
mean	26.000000	80.142857	212.714286	196.428571	333.142857	848.428571	11343.857143	0.002349	0.007461	0.019382	0.017186	0.030114	0.076493	5.0	5.0
std	15.491933	39.121301	89.697109	66.577917	51.725005	179.924852	2811.130107	0.001361	0.004293	0.008758	0.003979	0.004739	0.014672	0.0	0.0
min	0.000000	30.000000	60.000000	94.000000	255.000000	617.000000	7607.000000	0.000000	0.002392	0.005429	0.012357	0.023738	0.055832	5.0	5.0
25%	18.000000	46.500000	170.500000	161.000000	297.000000	733.500000	9868.000000	0.001631	0.004564	0.015611	0.014033	0.026967	0.067324	5.0	5.0
50%	30.000000	85.000000	212.000000	189.000000	355.000000	816.000000	11051.000000	0.002778	0.007248	0.019328	0.016489	0.029792	0.074591	5.0	5.0
75%	35.000000	116.500000	278.000000	244.000000	367.500000	943.500000	12228.500000	0.003230	0.009172	0.023282	0.020452	0.033049	0.086156	5.0	5.0
max	46.000000	120.000000	320.000000	282.000000	393.000000	1152.000000	16556.000000	0.003944	0.015118	0.033127	0.022486	0.037239	0.098068	5.0	5.0