# CapStone Project – The Battle of the Neighborhoods

## Vcare – Uplay

## Introduction

### Background & Problem

In a community with growing population of families and young children, the life of newborn and toddler Mother's is a challenge. In cases of working mother, full day at work and return to the second shift of continuous baby care with little help and no personal time. The other world of full-time mother raising postpartum depression in young mom's due to the unexpected challenges with continuous caring of the baby and very less personal time.

Like everyone else, it is particularly important for young mothers and mothers of small children to have personal time to relax, swim, day at the spa or physical activity at the fitness center. The challenge is to find the best suitable place for the babies to be engaged, happy and safe, childcare centers

The traditional childcare centers agenda is to target full day or half day clients monthly to secure regular income and continuous cash flow. These traditional childcare centers will work for office going parents whereas does not provide any comfort or weightage for stay-at-home mothers.

### Interest

In support of stay-at-home mothers and for the after-hour welfare of working moms, the childcare will have care takers to attend to babies and toddlers on hourly basis or a full day based on the requests. This modern approach to childcare when setup near fitness center/sports clubs/spa helps the mothers to have a break in their hectic day to care for themselves. This will promote health and emotional wellbeing of the mothers and the family.

The proposal of the modern childcare comes with various advantages from ownership perspective also. It can work with maximum two fulltime employees and rest by part timers. The centers can also provide discounted monthly packages for one or two hours for attraction and have tie-ups with the near by fitness centers/clubs frequented by mothers. With the improved technology, the owners should be able to manage volume of requests on weekdays versus weekends by having the customers block the slots online only. The greatest challenge would be to select the appropriate community to establish the modern daycare center in Toronto/Ottawa neighborhood of Ontario Canada. The client will be provided maximum of 4 options on studying both the neighborhoods. The tools and process to overcome the challenge will be elaborated in the Data section.

## Data Acquisition & Cleaning

The business requirement requires us to study the below parameters:

- Since we need to study all the neighborhoods to clearly understand the family background with the financial capability.

- We will need geographical coordinates for Ottawa and Toronto neighborhoods td understand proximity.
- Supportive business operations in the vicinity.

As step 1, we will use the postal codes to gather the Toronto & Ottawa Neighborhoods. Using the postal codes, the census data can be extracted from the Canadian government website to study the population statistics. The census data requires major data cleaning and screening to understand the impact of the variables.

The venues that can help with tie-up strategies are identified using the Foursquare API with the Neighborhood, latitude, and longitude details. The possible venue categories that can be approached for business collaboration shall include Playground, Children activity centers like academies, yoga studio, gym, pool, spa and sports club.

## Data Sources

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_K

https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=E

## Data Cleaning

### *Neighborhood Data*

The postal codes extraction from the webpage is done using the web scraping technique. The data could possibly contain postal codes without neighborhoods assigned. Those postal codes must be filtered from the extracted data and then the data frame is created with the neighborhood, postal code and borough information only as shown below and is completed for both Toronto and Ottawa neighborhoods.

Sample of Ottawa data is shown below:

| | Postalcode | Borough | Neighborhood |
|---|---|---|---|
| 0 | K2A | Ottawa | Highland Park, McKellar Park /Westboro /Glabar... |
| 1 | K4A | Ottawa | Fallingbrook |
| 2 | K1B | Ottawa | Blackburn Hamlet, Pine View, Sheffield Glen |
| 3 | K2B | Ottawa | Britannia /Whitehaven, Bayshore, Pinecrest |
| 4 | K4B | Ottawa | Navan |

Using the postal code, information the geospatial coordinates are added to the data frame. The coordinates have been extracted from the google maps and captured on a spreadsheet. The spreadsheet is transformed to a data frame and combined with the neighborhood data frame shown above:

| | Postalcode | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | K2A | Ottawa | Highland Park, McKellar Park /Westboro /Glabar... | 45.3884 | -75.7456 |
| 1 | K4A | Ottawa | Fallingbrook | 45.4738 | -75.4780 |
| 2 | K1B | Ottawa | Blackburn Hamlet, Pine View, Sheffield Glen | 45.4317 | -75.5645 |
| 3 | K2B | Ottawa | Britannia /Whitehaven, Bayshore, Pinecrest | 45.3483 | -75.8078 |
| 4 | K4B | Ottawa | Navan | 45.4210 | -75.4267 |

### *Venue Identification for collaborative business:*

Using the neighborhood name, latitude, longitude information and with the application of the Foursquare API, the venues with the below categories are extracted and tabulated:

- Playground
- Children activity centers like academies
- yoga studio
- gym
- pool
- spa
- sports club.

Sample data frame is as shown below:

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 11 | Regent Park, Harbourfront | 43.654260 | -79.360636 | Body Blitz Spa East | 43.654735 | -79.359874 | Spa |
| 42 | Regent Park, Harbourfront | 43.654260 | -79.360636 | St Lawrence Community Centre Pool | 43.649557 | -79.365063 | Gym Pool |

The venues are grouped by neighborhood and the venue counts are gathered and combined to the existing data frame.

### *Census Data Cleanup:*

Census data as the name indicates contains large amount of information namely, Population statistics, income statistics, income tax return statistics, household ownership statistics and the statistics with race and language.

On analyzing the dataset, we can extract the count of children from 0-4 years and the family size of the families residing in the neighborhood. The families could be both couple run family and single parent run families.

Since this daycare is targeted for people with optimum income, the income percentage of the neighborhood should be at least 70% or above. The general assumption is if the overall percentage is in general close to 95%, of which the employment income is the income ratio that is obtained from the dataset. The difference of 25% is said to be government transfers which is the support fund received from the government.

The extracted data will be as shown below:

| | Postalcode | ChildCount | CFamilySize | LPFamilySize | Income% |
|---|---|---|---|---|---|
| 0 | K2A | 710.0 | 3.9 | 2.5 | 65.1 |
| 1 | K4A | 3460.0 | 4.1 | 2.8 | 80.2 |

K1B.csv

Sample input file is

## Methodology

The daycare shall be profitable in communities based on two major categories: one the community should have relatively high employment rate and the second being, young families with good family size. Although there has been a huge dip in economy due to COVID-19, things are now going back to normal in many countries around the world and Canada is no exception to the case. The census data of 2016 from the Canadian government website is the most reliable data to study the possible family sizes and employment statistics.

Since there are many towns in the Toronto and Ottawa neighborhoods, we shall use the postal codes to from the Wikipedia website and gather the census data using the postal codes. The census data shall include the count of children in the neighborhood who are less than 4 years, family size run by couple and of single parents with the employment rate in the neighborhood. These four factors when analyzed shall provide the list of best possible towns for the possible establishment of childcare center.

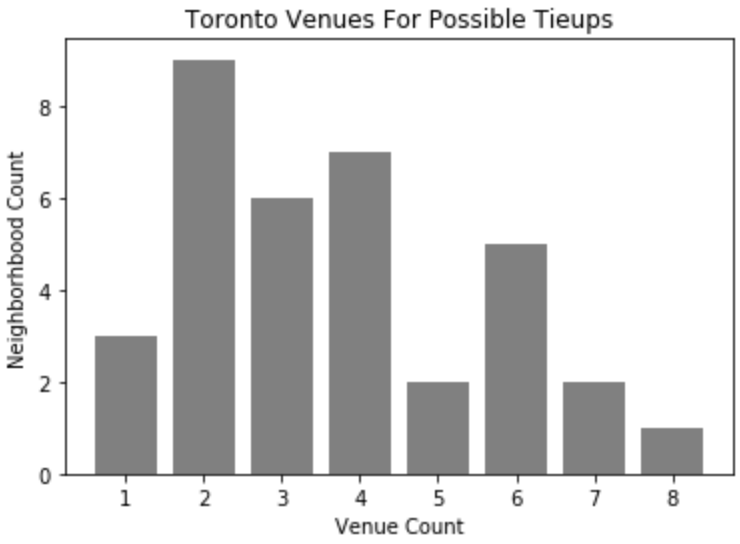
On the list of towns gathered, the Foursquare API is further applied to identify the location or presence of fitness center/sports clubs/spa to validate assurance of regular traffic and tie-up ventures to promote early growth of the center. Thereby the best locations shall be chosen for the client.

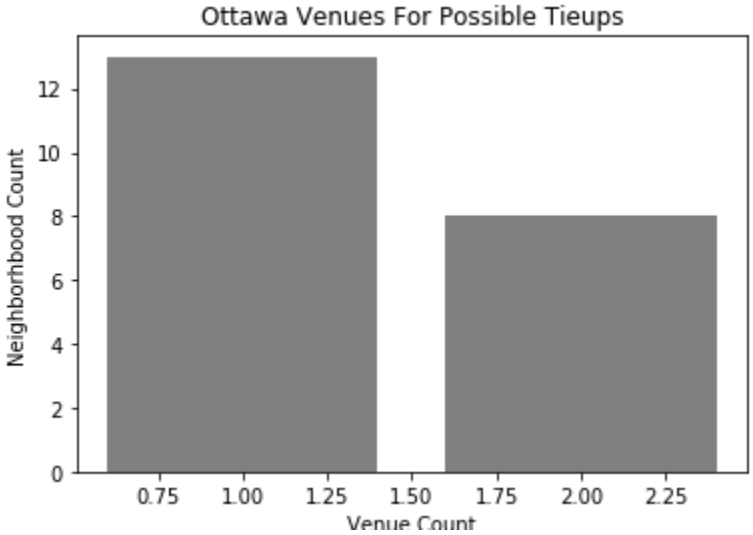
Since this a feature analysis by requirement, the machine learning algorithm to be applied will be K-means clustering. This is an unsupervised learning exercise and there are not many defined data groups or categories. The principal component analysis will be performed on the variables to assess the feature impact and the same will be used to plot the k-means cluster on a graph to determine the exact elbow point after which there is not much deviation in the cluster generation. This elbow point provides the optimum cluster number for the given data.

## Analysis

**Venue Count Analysis for the Neighborhoods**

Toronto and Ottawa grouped by Neighborhoods are as shown below. The graphs indicate that the venue for prospective business collaboration is more in Toronto neighborhood compared to Ottawa.

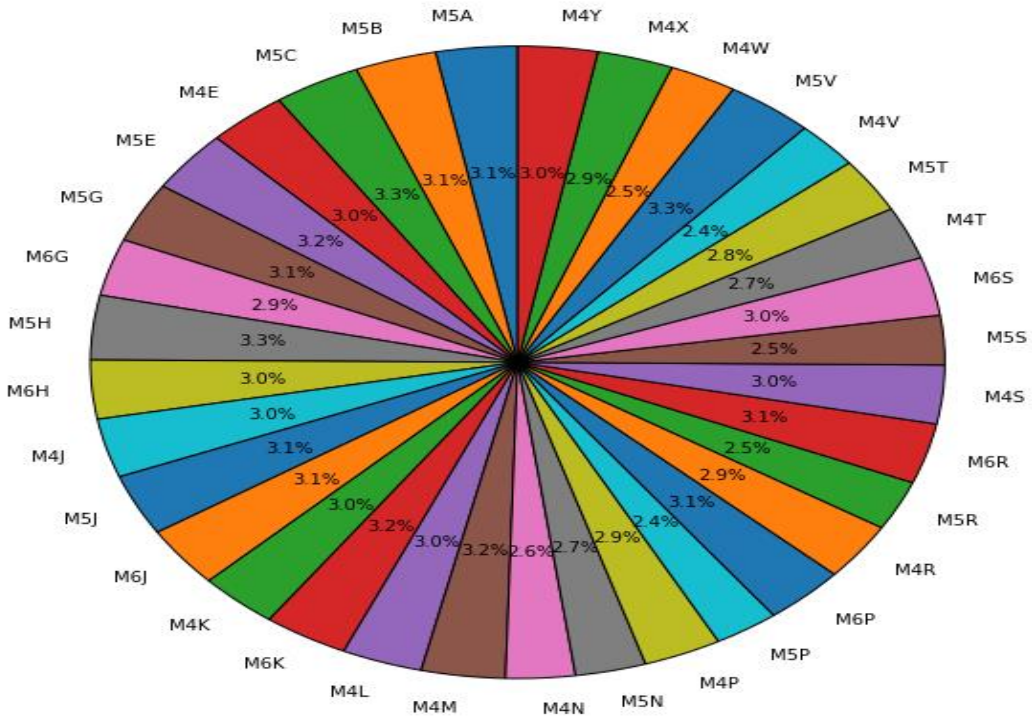


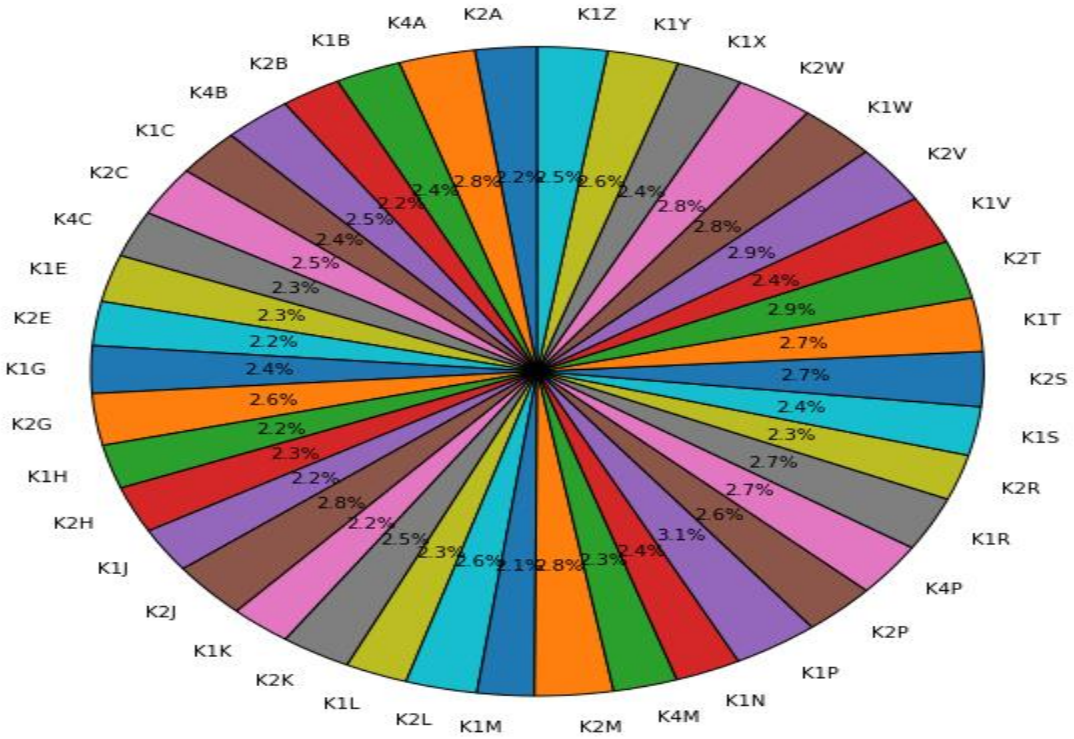

**Income Profile Analysis:**

The income profile analysis of the Toronto & Ottawa neighborhoods show that Toronto neighborhoods are more well-off. As we can see that pie charts, there are more neighborhoods/towns that have 3% and above as the income contribution over an 100% scale in Toronto. Whereas in Ottawa, all the towns except for one town have an income contribution of 3.1%

Toronto Borough Income Percentage Chart

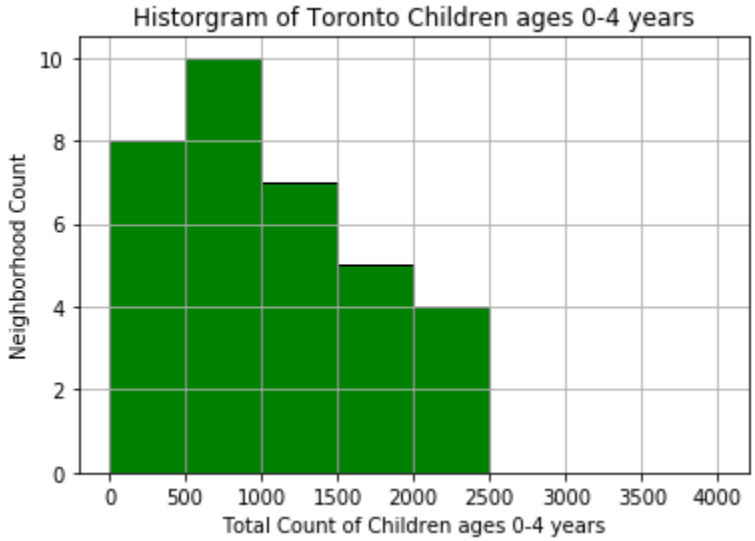


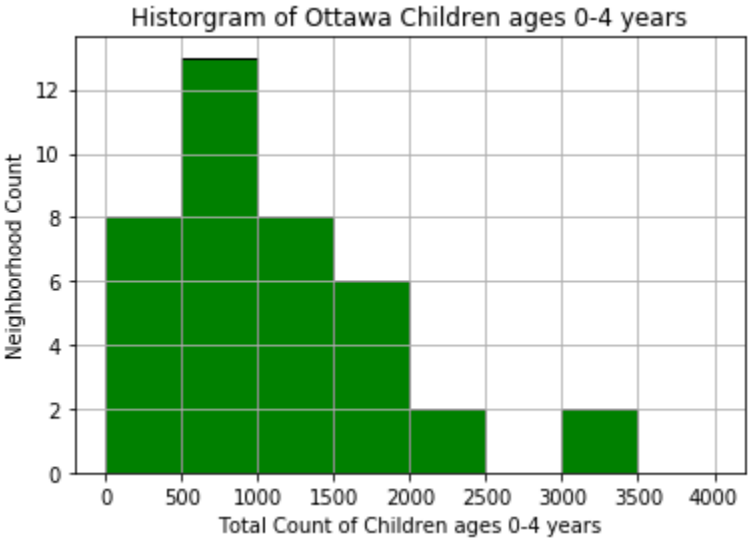

Ottawa Borough Income Percentage Chart



**Child Count Analysis:**

The child count analysis shows that the density of children relatively the same in Ottawa neighborhoods and the Toronto neighborhoods




## Machine Learning Application:

As we are using multiple variables, K-means clustering algorithm will be the best approach to find groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets

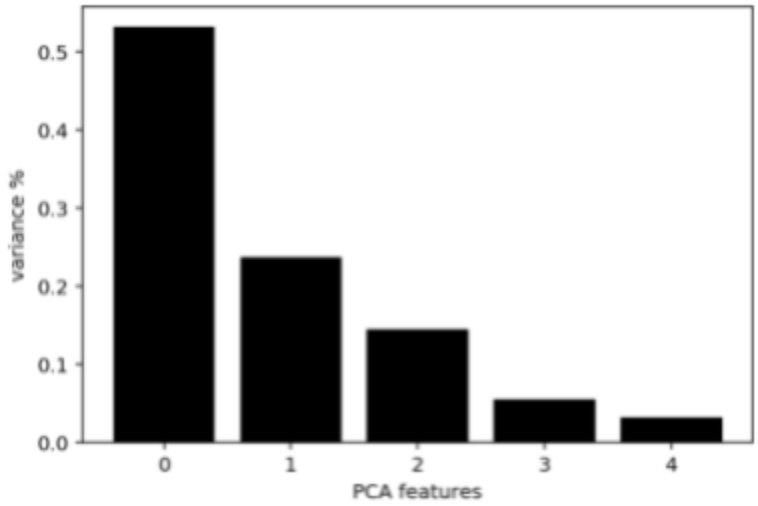
To achieve the results using K-means, all the three input data frames are combined to be subjected to the k-means cluster evaluation:

- Neighborhood data with geographical coordinates
- Venue Count by Neighborhood
- Census data by Neighborhood

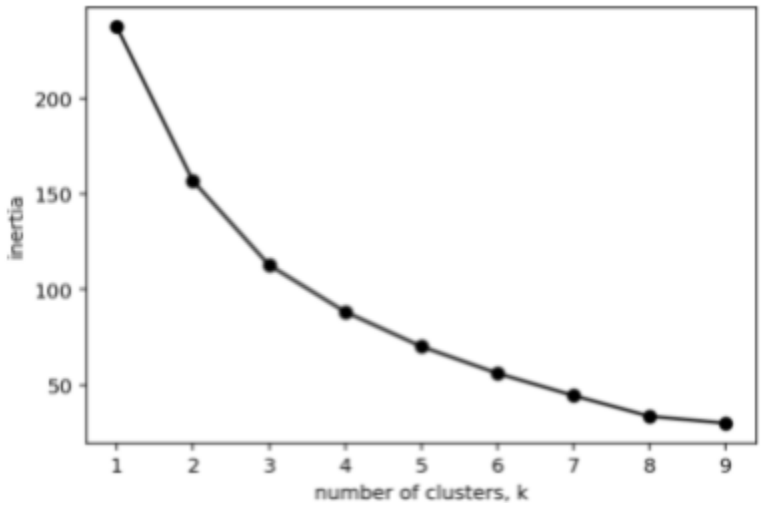
On performing the principal component analysis (PCA), we can determine from the below screenshot:

> Of the 5 features, we can see that Couple family size, Income % and Venuecount affect 90% of the variance ratio.

```
[0.53191579 0.23657486 0.14431731 0.0551395  0.03205255]
['CFamilySize', 'Income%', 'VenueCount', 'ChildCount', 'CFamilySize']
```




Using the PCA, the kmeans cluster is plotted on graph and analyzed against the measure of the sum of the squared distances to the nearest cluster center.




We can see that n=4 or 5 could be the most optimum as below n=5, the changes are very minimum. Using this result, the clusters have been generated for the dataset with cluster n=5 and the neighborhoods are selected from the best cluster that suits the condition.

## Results:

As per the cluster generation, the clusters observation can be elaborated as below:

Cluster #1: Cluster 1 has child count more than 3400 and venue count is only 1.

Cluster #2:  Cluster 2 has child count has less than 1000 and various venue count combination up to 8.

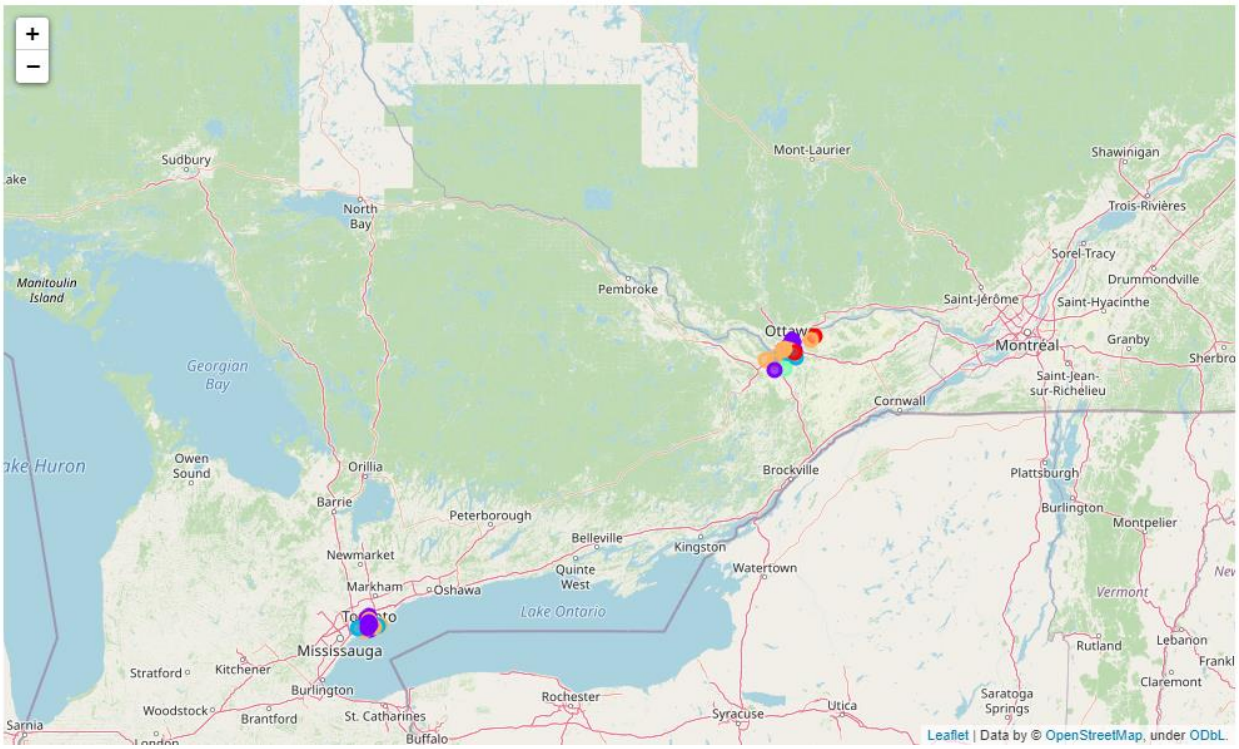
Cluster #3: Cluster 3 has child count more than 2000 and venue count combination up to 4.

Cluster #4: Cluster 4 has child count more than 4700 and venue count combination up to 1.

Cluster #5: Cluster 5 has child count between 1000 and 2000 and venue count combination up to 6.

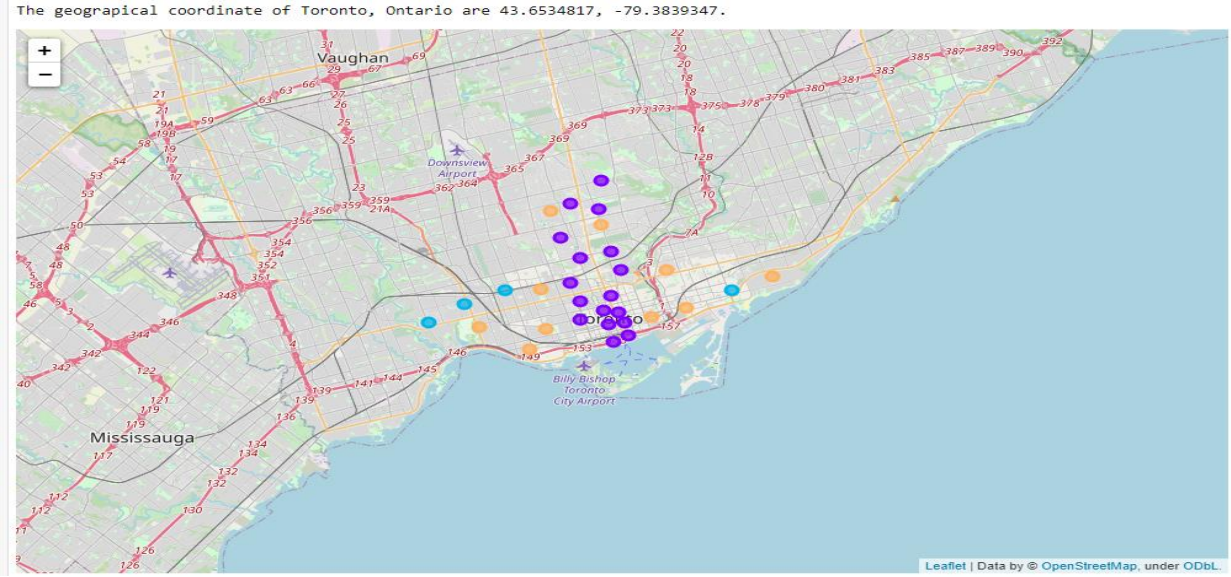
- Ontario Cluster Map showing both Toronto & Ottawa

The geograpical coordinate of Ontario, Canada are 50.000678, -86.000977.



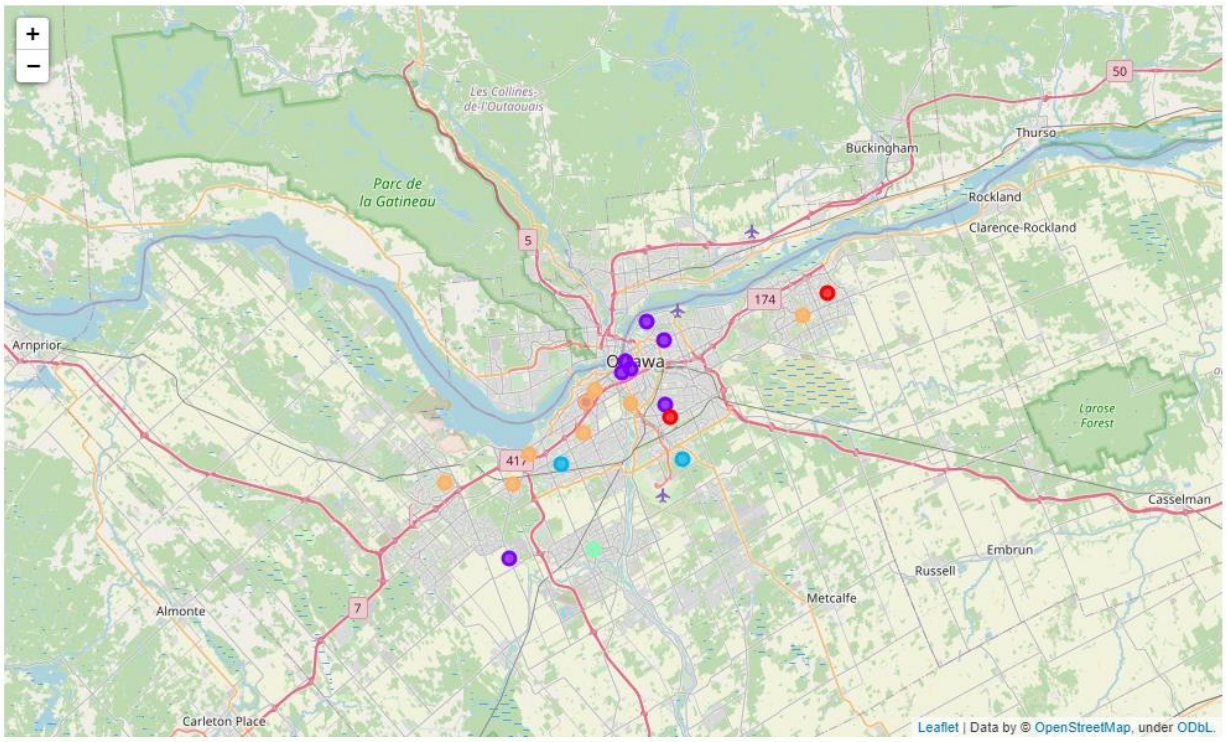

- Toronto Cluster Map

The geograpical coordinate of Toronto, Ontario are 43.6534817, -79.3839347.



- Ottawa Cluster Map

```
The geograpical coordinate of Ottawa, Ontario are 45.421106, -75.690308.
```




## Ideal Cluster selected:

| | Borough | Neighborhood | Latitude | Longitude | Income% | ChildCount | CFamilySize | LPFamilySize | VenueCount | ClusterLabels |
|---|---|---|---|---|---|---|---|---|---|---|
| **11** | Ottawa | Centrepointe, Meadowlands, City View, Craig He... | 45.340900 | -75.772500 | 74.6 | 2370.0 | 4.0 | 2.7 | 2.0 | 2 |
| **31** | Ottawa | Blossom Park, Greenboro, Leitrim, Findlay Creek | 45.344300 | -75.637600 | 77.1 | 2400.0 | 4.2 | 2.9 | 1.0 | 2 |
| **50** | West Toronto | Dufferin, Dovercourt Village | 43.669005 | -79.442259 | 80.2 | 1970.0 | 3.9 | 2.6 | 2.0 | 2 |
| **75** | East Toronto | India Bazaar, The Beaches West | 43.668999 | -79.315572 | 81.3 | 2260.0 | 3.9 | 2.6 | 2.0 | 2 |
| **104** | West Toronto | High Park, The Junction South | 43.661608 | -79.464763 | 81.9 | 2210.0 | 3.8 | 2.5 | 4.0 | 2 |
| **122** | West Toronto | Runnymede, Swansea | 43.651571 | -79.484450 | 79.5 | 2175.0 | 3.8 | 2.6 | 3.0 | 2 |

# Conclusion:

The original goal of the project was to identify the neighborhoods in Toronto and Ottawa borough where they can establish the modern daycare center servicing only ages 0-4 years. with families having optimum income and have venue options to support business tie-ups. Of cluster# 3, the best suited neighborhood for the daycare is HighPark, The Junction South in East Toronto and for Ottawa borough, the option will be Centrepointe, Meadowlands, City View, Craig Henry, Tangelwood, Grenfell Glen, Davidson Heights.

- The count of children 0-4 age years is >=2000
- The income reange is >=75%
- The venue count for business tie-ups is also available.