SOFE 3720U: Introduction to Artificial Intelligence

Intro. to Artificial Intelligence Project

Group #1

Team members' names and ID

Name 1: Foram Gandhi	**ID 1:** 100699245
Name 2: Kinjal Shah	**ID 2:** 100743551
Name 3: Rutvi Shah	**ID 3:** 100747171

Professor: Dr. Masoud Makrehchi

Date: April 11th, 2022

Introduction: Background and Business Problem

Toronto is the largest city in Ontario, with a population of over 2.7 million people. With a high population that is growing day-by-day, the level of crime is bound to increase. Toronto's vibrant culture, convenience and educational environments are a few of many reasons that families decide to move to the city. Parents strive to find safe and accessible places to live so they can raise their current/future children. **Thus, the goal of this project is to find the best neighbourhood in Toronto to raise children**. This takes three factors into account:

1. **Nearby schooling options** (preschool, primary, elementary, middle, high, and secondary schools)
2. **Crime rate** (Break and Enter)
3. **Children population** (between the ages 0-14)

In order to achieve the goal, neighbourhoods of the City of Toronto were segmented into clusters and the correlation between their relative crime rates and population of children in the age group of 0 to 14 years old was analyzed. We utilized data from the Toronto Neighbourhood dataset, Neighbourhood Crime Rates dataset, Toronto Neighbourhood Profiles dataset and Foursquare's Places API. This data, combined with the use of an unsupervised machine learning algorithm called K-means Clustering, further aided us in our analysis. We hope that the results of this project can help families make a more informed decision about which neighbourhood they want to raise their children in.

Data Explanation and Data Sources

The first data source used, as mentioned previously, was the Neighbourhoods dataset, from which we obtained the names of all neighbourhoods present in Toronto, as well as a GeoJSON file of all the areas. This allowed us to create an accurate representation of the map using a Python library called Folium, by taking the latitude and longitude metrics into account (Fig. 1).

Source: https://open.toronto.ca/dataset/neighbourhoods/

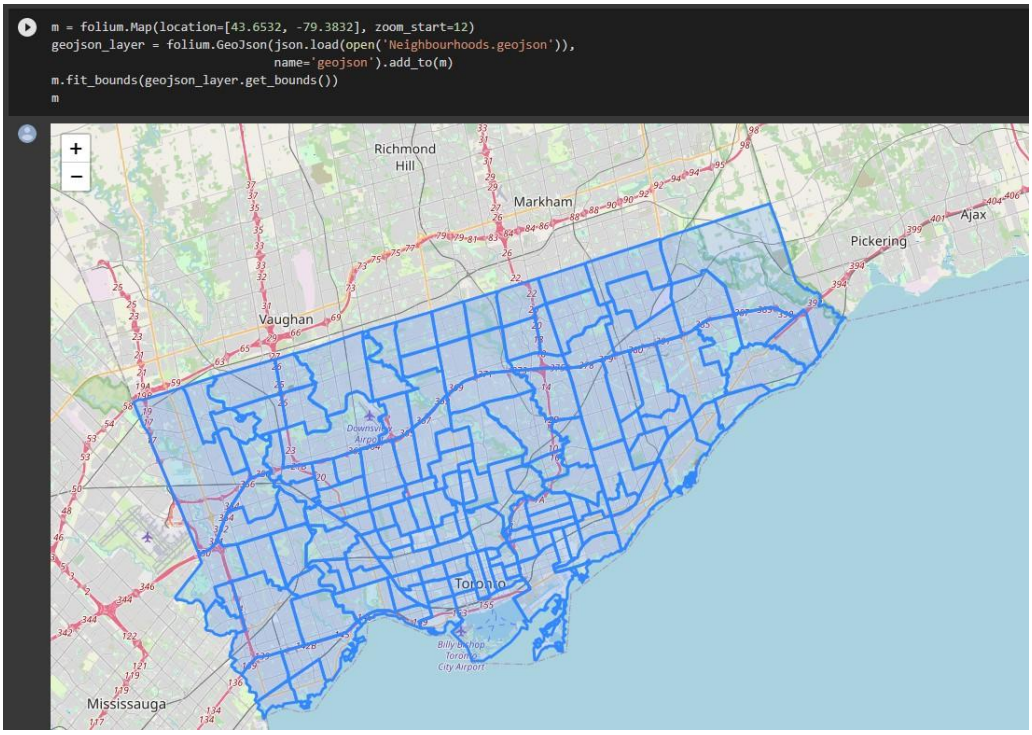

Fig. 1: Using Folium to map out the neighbourhoods from the GeoJSON file

The third data source used was the Neighbourhood Crime Rates dataset. The “Break and Enter” information for the year of 2016 was acquired and placed into a .CSV file. This data was an important factor in determining the attractiveness of a neighbourhood. A higher crime rate meant a lower chance of a family wanting to move to the location. The focus was drawn on the “Break and Enter” category because oftentimes, parents will leave their children home alone while they are at work, thus it is likely that they would desire a lower break and enter rate in their neighbourhood.

Source:
https://data.torontopolice.on.ca/datasets/neighbourhood-crime-rates-boundary-file-/explore

Foursquare’s Places API was the second data source used in order to fetch information about nearby schooling options. Foursquare is an independent location data platform that offers real-world data of places and venues all over the world. Our project consisted of filtering the query for venue name to “schools” and the near query to the specific location of the neighbourhood, for example, Agincourt, Toronto” (Fig. 2). In order to make actual requests to the API for data, an authentication token was obtained.

Source: https://developer.foursquare.com/reference/place-search

QUERY PARAMS

query string

A string to be matched against all content for this place, including but not limited to venue name, category, taste, and tips.

schools

near string

A string naming a locality in the world (e.g., "Chicago, IL"). If the value is not geocodable, returns an error. One of ll, ne and sw, or near must be specified.

Agincourt, Toronto

Fig. 2: Initial experimentation of an API search query


```
{
  "results": [
    {
      "fsq_id": "4baa8ebcf964a5206b743ae3",
      "categories": [
        {
          "id": 12059,
          "name": "High School",
          "icon": {
            "prefix": "https://ss3.4sqi.net/img/categ
            "suffix": ".png"
          }
        }
      ],
      "chains": [],
      "distance": 1206,
      "geocodes": {
        "main": {
          "latitude": 43.788818,
          "longitude": -79.278925
        },
        "roof": {
          "latitude": 43.788818,
          "longitude": -79.278925
        }
      },
      "link": "/v3/places/4baa8ebcf964a5206b743ae3",
      "location": {
        "address": "2621 Midland Ave",
        "country": "CA",
        "cross_street": "",
        "formatted_address": "2621 Midland Ave, Scarb
        "locality": "Scarborough",
        "neighborhood": [
          "Agincourt"
        ],
        "postcode": "M1S 1R6",
        "region": "ON"
      },
      "name": "Agincourt Collegiate Institute",
      "related_places": {},
      "timezone": "America/Toronto"
    },
```


Fig. 3: Response from the example search query

The fourth data source utilized was the Neighbourhood Profiles dataset, which provided more detailed information about the children population in each neighbourhood. We considered the age range of 0-14 years, as the target audience for this project would be adults planning on having kids soon and adults with children.

Source: https://open.toronto.ca/dataset/neighbourhood-profiles/

Methodology

The Foursquare Places API and various datasets were explored to determine the best choice of use for our project's goal. After deciding on them, the pertaining data was extracted from each .csv file and placed into a new .csv file for further analysis.

As shown in Fig. 4 below, a call to the Foursquare Places API was made to fetch all the 'school' venues near a given neighbourhood. The response was then parsed to find all the venue names that had certain keywords in the results, such as "elementary", "secondary", and "middle". If so, the number of schools in the corresponding neighbourhood was incremented. Without the parsing step, the response would have also included venues like driving schools, art schools, and generally any venues with the word 'school' in their name. Therefore, this step ensured that we only considered educational institutions within the neighboruhoods.

```
headers = {
    "Accept": "application/json",
    "Authorization": AUTHORIZATION
}

df = pd.read_csv('Age_Vs_Crime.csv')
neighborhoods = df.columns[:]
school_count = [0]*140

for n in range(len(neighborhoods)):
    url = "https://api.foursquare.com/v3/places/search?query=school&near={}%2C%20Toronto".format(neighborhoods[n])
    response = requests.request("GET", url, headers=headers).json()

    names = ["elementary", "primary", "secondary", "high", "middle", "preschool"]

    for i in response["results"]:
        for j in i["categories"]:
            if any(name in j["name"].lower() for name in names):
                school_count[n] += 1
print(school_count)
```

```
[5, 6, 4, 2, 7, 2, 1, 1, 5, 4, 7, 6, 3, 1, 3, 3, 8, 2, 1, 0, 7, 0, 0, 7, 1, 7, 2, 0, 7, 7, 5, 2, 7, 7, 2, 7, 2, 1, 8, 3, 5, 0, 2, 1, 4,
```

Fig. 4: Shows the algorithm to retrieve the number of schools in each area and the count array printed

Fig. 5 shows a table with all the information that was gathered from the multiple data sources, which was then placed into the .csv file and pulled into a table format.

```
df = pd.read_csv('Age_Vs_Crime.csv')
df = df.set_index([pd.Index(['Pop. of Ages 0-14', 'Crime Rate of B&E', 'Num. of Schools'])])
df = df.T
df['Pop. of Ages 0-14'] = df['Pop. of Ages 0-14'].astype(str).str.replace(',', '').astype(float)
df['Crime Rate of B&E'] = df['Crime Rate of B&E'].astype(str).str.replace(',', '').astype(float)
df['Num. of Schools'] = df['Num. of Schools'].astype(str).str.replace(',', '').astype(float)
df
```

	Pop. of Ages 0-14	Crime Rate of B&E	Num. of Schools
Agincourt North	3840.0	49.0	5.0
Agincourt South-Malvern West	3075.0	61.0	6.0
Alderwood	1760.0	26.0	4.0
Annex	2360.0	123.0	2.0
Banbury-Don Mills	3605.0	57.0	7.0
...	...	...	...
Wychwood	1860.0	29.0	7.0
Yonge-Eglinton	1800.0	19.0	7.0
Yonge-St.Clair	1210.0	12.0	7.0
York University Heights	4045.0	98.0	0.0
Yorkdale-Glen Park	1960.0	66.0	7.0

140 rows × 3 columns

Fig. 5: Shows the information from the CSV file put into a table

Clustering is a method used for grouping a collection of data points that have similarities together. The K-means clustering algorithm is an unsupervised ML algorithm that helps discover patterns within data. It defines a fixed number of clusters (k) in a dataset. For this project, we chose k=5 clusters.

One of the methods used to evaluate the clusters is the Elbow Method (Fig. 6). It determines the number of clusters in the data set by calculating the sum of the samples' squared distances. Upon plotting the graph, a point of inflection can be observed, which is known as the elbow point. We noticed that at k=5, there was no outstanding decrease in the sum of square distances, thus it was the elbow point and we would be creating 5 clusters for our data.

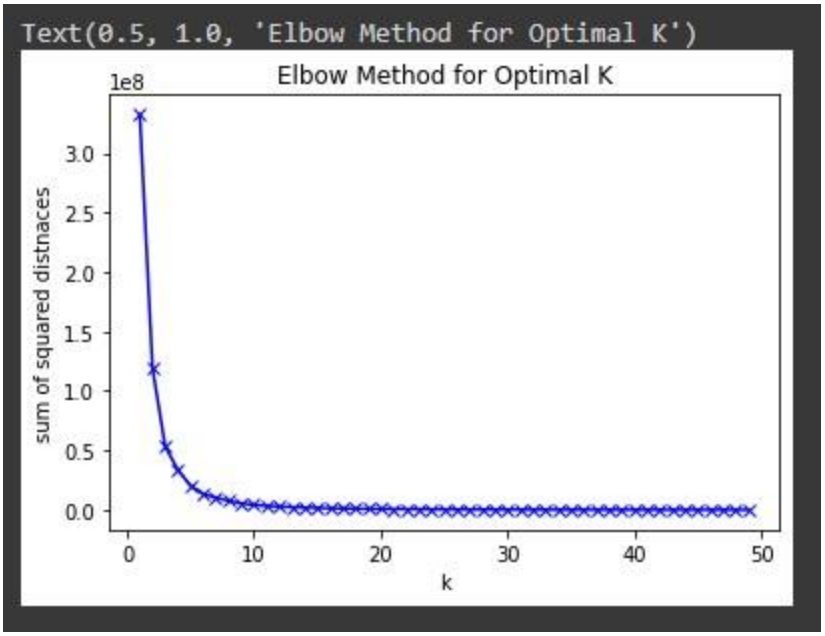

Fig. 6: The K-means Elbow Method plotted on a graph

Another method we used to find the optimal k was the Silhouette Method (Fig. 7). To measure the efficiency of a clustering technique, a metric called the silhouette coefficient was utilized. There were two factors that the points being assigned into clusters were based on - closeness of data points in the same cluster (cohesion) and distance between the data points in other clusters (separation). The average silhouette coefficient was summarized into one score. The score() function required a minimum of two clusters in order to run without exceptions. A list was used to store the coefficients for each k and the range went from 2 to 49. As shown in Fig. 6, there is a distinct peak at k=5, therefore, that would be the chosen k-value for the number of clusters.

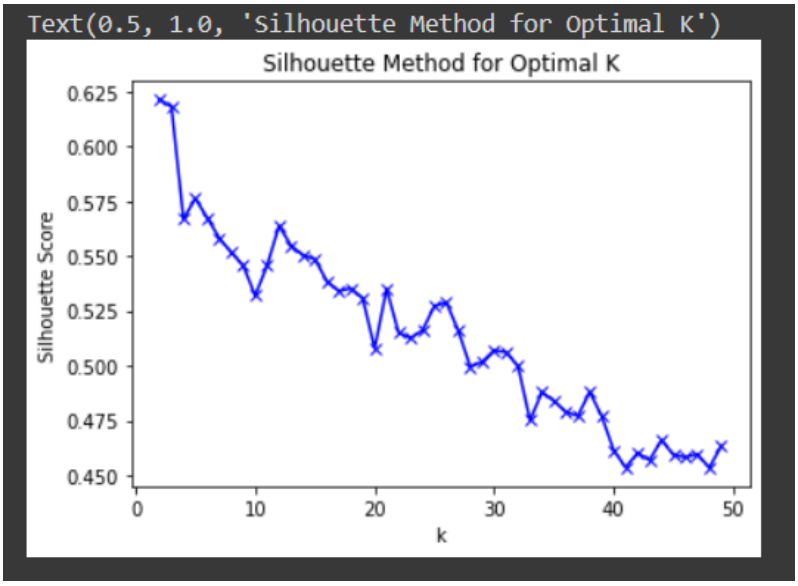

Fig. 7: The K-means Silhouette Method plotted on a graph

The method that granted more of an optimal k-value, in this case, was the Elbow Method. This method was chosen because it had more of a distinct drop in comparison to the Silhouette Score. Both were graphed in order to identify whether the same or different results were given in terms of the optimal k-value. Since the Elbow Method was selected, the k-value would remain as 5.

Clustering

After getting the optimal k value of 5, the sklearn.cluster library was used to perform K-Means clustering on the data frame that included the population of children between the ages 0-14, the crime rate for B&E and the number of schools in each neighbourhood (Refer to Fig. 5). The results of the clustering can be seen in Fig. 8 and Fig. 9 displaying the number of neighbourhoods in each cluster and a data frame displaying the neighbourhood associated with the cluster number, respectively.

```
#number of clusters
k = 5
#run K-Means clustering algorithm
cluster = KMeans(n_clusters=k).fit(df)
#prints the counts of neighbourhoods assigned to diff. clusters
print(collections.Counter(cluster.labels_))
```

```
Counter({1: 62, 0: 38, 4: 28, 2: 8, 3: 4})
```

Fig. 8: Performed the K-Means Clustering with 5 clusters fit on the data frame from Fig. 5. The number of neighbourhoods in each cluster is then printed. There are 62 neighbourhoods assigned to cluster 1, 38 to cluster 0, 28 to cluster 4, 8 to cluster 2 and 4 to cluster 3.

```
#labels
labels = cluster.labels_
results = pd.DataFrame([df.index, labels]).T
results
```

	0	1
0	Agincourt North	4
1	Agincourt South-Malvern West	0
2	Alderwood	1
3	Annex	0
4	Banbury-Don Mills	4
...	...	...
135	Wychwood	1
136	Yonge-Eglinton	1
137	Yonge-St.Clair	1
138	York University Heights	4
139	Yorkdale-Glen Park	1

140 rows × 2 columns

Fig. 9: Results of the clustering

Scoring

The scoring of each neighbourhood was determined by using the formula below:

$$\frac{[Pop.\ of\ Ages\ 0{-}14]}{[Crime\ Rate\ of\ B\&E]} \times [Num.\ of\ Schools]$$

We chose to proceed with this formula because as the goal states, we want to find the best neighbourhood to raise children. Therefore, the population of children and the number of schools should have a positive effect on the score (increase score → numerator) while crime rate should have a negative effect (decrease score → denominator).

	Pop. of Ages 0-14	Crime Rate of B&E	Num. of Schools	Score	Cluster
Agincourt North	3840.0	49.0	5.0	391.836735	4
Agincourt South-Malvern West	3075.0	61.0	6.0	302.459016	0
Alderwood	1760.0	26.0	4.0	270.769231	1
Annex	2360.0	123.0	2.0	38.373984	0
Banbury-Don Mills	3605.0	57.0	7.0	442.719298	4
...	...	...	...	...	...
Wychwood	1860.0	29.0	7.0	448.965517	1
Yonge-Eglinton	1800.0	19.0	7.0	663.157895	1
Yonge-St.Clair	1210.0	12.0	7.0	705.833333	1
York University Heights	4045.0	98.0	0.0	0.000000	4
Yorkdale-Glen Park	1960.0	66.0	7.0	207.878788	1

140 rows × 5 columns

Fig. 10: The data frame including the cluster number and calculated score for each neighbourhood

Results

Fig. 11-14 represent the number of neighbourhoods in each cluster.

	Pop. of Ages 0-14	Crime Rate of B&E	Num. of Schools	Score	Cluster
2	1760.0	26.0	4.0	270.769231	1
6	1695.0	120.0	1.0	14.125000	1
8	1515.0	33.0	5.0	229.545455	1
10	1120.0	11.0	7.0	712.727273	1
14	1405.0	23.0	3.0	183.260870	1
...	...	...	...	...	...
134	1165.0	12.0	7.0	679.583333	1
135	1860.0	29.0	7.0	448.965517	1
136	1800.0	19.0	7.0	663.157895	1
137	1210.0	12.0	7.0	705.833333	1
139	1960.0	66.0	7.0	207.878788	1

62 rows × 5 columns

Fig. 11: Number of neighbourhoods in cluster 1

	Pop. of Ages 0-14	Crime Rate of B&E	Num. of Schools	Score	Cluster
33	5725.0	52.0	7.0	770.673077	2
45	5790.0	30.0	7.0	1351.000000	2
58	5820.0	105.0	2.0	110.857143	2
65	6120.0	9.0	6.0	4080.000000	2
95	5840.0	46.0	3.0	380.869565	2
118	5600.0	24.0	0.0	0.000000	2
124	5060.0	131.0	0.0	0.000000	2
129	5920.0	78.0	7.0	531.282051	2

Fig. 12: Number of neighbourhoods in cluster 2

	Pop. of Ages 0-14	Crime Rate of B&E	Num. of Schools	Score	Cluster
73	7910.0	65.0	5.0	608.461538	3
81	7090.0	34.0	7.0	1459.705882	3
105	7960.0	75.0	1.0	106.133333	3
132	9625.0	80.0	2.0	240.625000	3

Fig. 13: The number of neighbourhoods in Cluster 3

	Pop. of Ages 0-14	Crime Rate of B&E	Num. of Schools	Score	Cluster
0	3840.0	49.0	5.0	391.836735	4
4	3605.0	57.0	7.0	442.719298	4
9	4555.0	97.0	4.0	187.835052	4
11	4550.0	57.0	6.0	478.947368	4
13	4600.0	27.0	1.0	170.370370	4
24	4565.0	86.0	1.0	53.081395	4
30	4065.0	65.0	5.0	312.692308	4
31	4250.0	70.0	2.0	121.428571	4
32	4480.0	78.0	7.0	402.051282	4
35	3735.0	77.0	7.0	339.545455	4
37	4180.0	65.0	1.0	64.307692	4
39	4805.0	59.0	3.0	244.322034	4
42	4385.0	14.0	2.0	626.428571	4
50	3540.0	36.0	1.0	98.333333	4
63	4240.0	34.0	2.0	249.411765	4
76	3675.0	67.0	7.0	383.955224	4
77	3505.0	48.0	0.0	0.000000	4
102	3505.0	38.0	7.0	645.657895	4
110	3975.0	148.0	5.0	134.290541	4
113	4060.0	49.0	3.0	248.571429	4
114	3885.0	61.0	4.0	254.754098	4
116	3825.0	77.0	3.0	149.025974	4
122	3650.0	212.0	7.0	120.518868	4
123	4635.0	72.0	0.0	0.000000	4
125	3440.0	20.0	4.0	688.000000	4
128	4215.0	76.0	7.0	388.223684	4
131	3555.0	35.0	2.0	203.142857	4
138	4045.0	98.0	0.0	0.000000	4

Fig. 14: The number of neighbourhoods in Cluster 4

The image below (Fig. 15) displays the map of Toronto neighbourhoods colour-coded and segmented into their respective cluster groups. When hovering over a neighbourhood, its name and score can be observed.

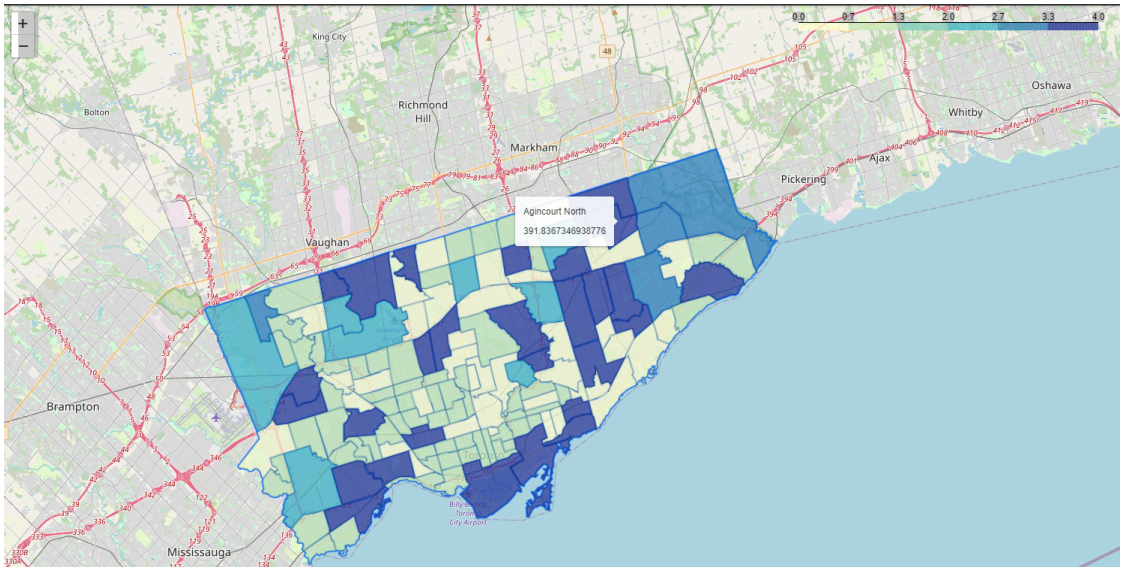

Fig. 15: Map of neighbourhood clusters with scores

To find the best cluster for the goal aforementioned, the mean score of each cluster was calculated (Fig. 16). According to the results, cluster 2 has the highest mean score, followed by clusters 3, 0, 4 and 1.

```
Cluster
0    320.653869
1    247.745093
2    903.085230
3    603.731439
4    264.266136
Name: Score, dtype: float64
```

Fig. 16: Cluster numbers and their calculated mean scores.

Discussion and Conclusion

The best neighbourhood for a family with children between the ages of 0 and 14 would be any neighbourhood in cluster 2, such as L'Amoreaux or Glenfield-Jane Heights, because the neighbourhoods in that cluster are deemed to be the safest in terms of the average score of that cluster (Refer to Fig. 16). Although some neighbourhoods, such as West Humber-Clairville, have a score of 0, this simply means that there were no schools in that area. Cluster 2 is preferable because the mean score of 903 (approximately) is the highest. Cluster 3 comes to a close second with a mean score of about 603, however, compared to cluster 2 the populations in cluster 3 are

greater. The crime rate is also greater, which in turn lowers the score of the neighbourhoods in cluster 3.

A correlation between crime rate and population of children in specific Toronto neighbourhoods has been established by first looking at 3 factors - nearby schooling options, crime rate and children population within the ages of 0-14. Next, the data samples were mapped into graphs and clustered using the K-Means clustering algorithm. Finally, analysis on the clusters was performed. According to our findings, our recommendation for a family looking to move to Toronto would be to reside in the L'amoreaux neighbourhood as it has the highest score (4080) in the best cluster (cluster 2).

The findings for this project can be improved if a more current dataset is utilized, instead of looking at limited past data. Also, APIs that provide more venue-specific data (focusing on educational institutions) can help provide more accurate results.