Finding a good neighborhood for your family in the city of Toronto

Introduction:

People with kids, coming from suburbs, small towns or other countries to Toronto city looking for an accommodation have one common problem i.e. finding a good neighborhood for a home. Their first criterion is having a good school in the neighborhood. And then other amenities like Restaurants and Parks etc.

This problem is faced by families all over the world to varying extents. Most impacted are the families with kids who have to be more aware of the neighborhood. In some areas, school is the primary driver while in others public transport may be the primary driver.

So this project is intended to guide one to solve this type of problem. Depending on the location and the influencing drivers this project can be modified.

Data:

The dataset will be derived from different sources:

--> Location data for Toronto will be used from the previous assignments where we got the postal codes, latitudes and longitudes for Toronto city.

	Postalcode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

--> Venue data will be extracted using Foursquare API's using the location latitudes and longitudes.

--> School data will be extracted from a website. Initial idea was to scrape the HTML data from this website for all postal codes. But this site provides postal code only as an internal filter without any change in the URL. So finally the approach taken was to download data for each postal code one at a time and then merge all the postal codes in one file.
http://ontario.compareschoolrankings.org/elementary/SchoolsByRankLocationName.aspx?schooltype=elementary

	Rank	Rating	Postalcode	School
0	165/3064	8.4	M4E	Balmy Beach
1	224/3064	8.2	M4E	St Denis
2	567/3064	7.5	M4E	St John
3	770/3064	7.2	M4E	Williamson Road
4	986/3064	6.9	M4E	Adam Beck

Methodology:

The aim was to find few of the important venues in a neighborhood that can impact a home buyer or renter, who in our case is a family with kids. I didn't consider population, size of neighborhood, crime rate etc. for this project.

Data analysis was done to understand:
--> What type of venues should be considered? For families with kids, school is a very important criteria. Next I checked for Hospitals and Parks using Foursquare and found that there are a few around most neighborhoods. So I narrowed down my search to Grocery Shops and Restaurants where data was more random.

--> For schools, using foursquare I could not get the full list of elementary schools along with their ratings in Toronto. So I had to use a website to download data for schools by postal code.

--> I wanted to include count as well as the mean rating for restaurants and grocery shops. Using Foursquare I checked for a random sample, and found that most of them don't have a rating on Foursquare. So I chose to only include only counts for Restaurant and Park in the dataset using Foursquare.

I used kmeans clustering to group neighborhoods based on these criteria. School Ratings, Park counts, Restaurant counts were passed to the clustering algorithm to give the results in the form of 4 different clusters. This algorithm was best suited for this kind of clustering problem as per the dataset.

```
toronto_parks_count.head()
```

	Neighborhood	id	name	categories
0	Adelaide, King, Richmond	4	4	4
1	Berczy Park	4	4	4
2	Brockton, Exhibition Place, Parkdale Village	5	5	5
3	Business reply mail Processing Centre969 Eastern	1	1	1
4	Cabbagetown, St. James Town	3	3	3

```
toronto_restaurant_count.head()
```

	Neighborhood	id	name	categories
0	Adelaide, King, Richmond	30	30	30
1	Berczy Park	14	14	14
2	Brockton, Exhibition Place, Parkdale Village	4	4	4
3	Cabbagetown, St. James Town	9	9	8
4	Central Bay Street	30	30	30

Results:

The idea is to group neighborhoods that have similar characteristics i.e. schools followed by grocery shops and restaurants. There are 38 neighborhoods in Toronto containing Toronto in the Borough name. On exploring venues using Foursquare, I see that there are 34 neighborhoods with Parks and 30 neighborhoods with Restaurants in a 500 meter radius around the neighborhood location. For Parks and Restaurants, I group by neighborhood to get the counts.

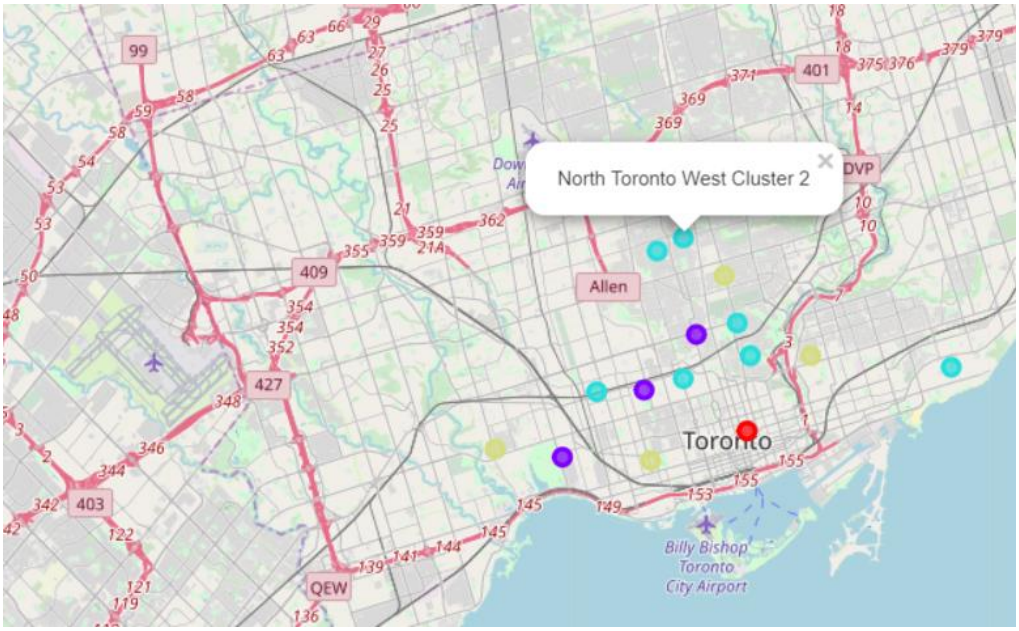
Since the first criteria is School, we narrow down our list to the neighborhoods having good elementary schools. For this I check elementary school in all 38 neighborhoods using the school data and find 29 elementary schools with rating >= 8. Next I group these schools by neighborhood to get the mean rating. Finally I have 16 neighborhoods with good schools to looks at.

Now I merge these 16 neighborhoods with the Parks and Restaurants dataset to arrive at our final data for clustering. Among these 16 neighborhoods, I have few with either no parks or no restaurants.

Finally I run kmeans clustering on this final dataset with a cluster size of 4. The kmeans clustering algorithm does a good job in clustering the neighborhoods based on schools, parks and restaurants. The results show that majority of neighborhoods i.e. 8 out of 16 have the same characteristics. They have good schools, few parks and very few restaurants. The next two big clusters have 3 neighborhoods each. School ratings are almost the same whereas one has fewer parks and restaurants than the other. The smallest cluster has only one neighborhood with a very high school rating and a lot of restaurants. Based on these clusters, one can recommend a neighborhood to a family or let the family choose a neighborhood based on this result set.

```
df.sort_values('Cluster Labels') # check the last columns!
```

	Neighborhood	Latitude	Longitude	SchoolRating	ParksCount	RestaurantCount	Cluster Labels
12	Ryerson, Garden District	43.657162	-79.378937	9.900000	3.0	30.0	0
0	Christie	43.669542	-79.422564	8.200000	2.0	3.0	1
2	Deer Park, Forest Hill SE, Rathnelly, South Hi...	43.686412	-79.400049	9.800000	1.0	4.0	1
8	Parkdale, Roncesvalles	43.648960	-79.456325	8.200000	0.0	4.0	1
3	Dovercourt Village, Dufferin	43.669005	-79.442259	10.000000	6.0	1.0	2
4	Lawrence Park	43.728020	-79.388790	9.100000	1.0	0.0	2
6	Moore Park, Summerhill East	43.689574	-79.383160	8.566667	2.0	0.0	2
7	North Toronto West	43.715383	-79.405678	9.300000	2.0	2.0	2
9	Rosedale	43.679563	-79.377529	8.700000	3.0	0.0	2
10	Roselawn	43.711695	-79.416936	8.450000	0.0	0.0	2
13	The Annex, North Midtown, Yorkville	43.672710	-79.405678	9.300000	2.0	0.0	2
14	The Beaches	43.676357	-79.293031	8.300000	3.0	1.0	2
1	Davisville	43.704324	-79.388790	8.200000	3.0	6.0	3
5	Little Portugal, Trinity	43.647927	-79.419750	8.400000	4.0	8.0	3
11	Runnymede, Swansea	43.651571	-79.484450	8.766667	3.0	7.0	3
15	The Danforth West, Riverdale	43.679557	-79.352188	8.900000	3.0	9.0	3



Conclusion:

Adding more attributes like restaurant rating would have been more useful for the end user to be able to select neighborhoods with good restaurants. Since Foursquare doesn't have good ratings data in this case, would have ideally liked to use Google or other API's instead.

Observing the clusters, one can choose which neighborhoods to look at for finding a home.
You can use this process and algorithm in other use cases as well. Ex. to find a good neighborhood for a new restaurant or a new grocery shop. Or find a neighborhood with low crime rate and population.

The major takeaway from this project is that one can customize this dataset to add more venues as well as more attributes. For ex, we can chose to add Grocery shops or Hospitals to the venue list. Similarly one can choose to include the Restaurant rating along with the count for each neighborhood. So one can have the results based on their set of criteria.