Expatriated executive to Toronto City looking for a new family home

This is the last part of the IBM Capstone Project.

Objectives of the final assignments were to define a business problem, look for data in websites and use Foursquare location data.

1. Discussion and Background of the Business Problem:

Problem Statement: **An executive manager has been expatriated to Toronto City and has to look for a new family home in one of its neighbourhoods.**

The objective is to evaluate the selection of a suitable neighbourhood in a city based on key criteria predefined.

In this case the target city is Toronto, the largest in Canada, and one of the largest in North America (behind only Mexico City, New York and Los Angeles). With a population just short of 3 million people, but The Greater Toronto Area (GTA) includes around 6.5 million people, stretching along the shore of Lake Ontario and including suburban communities further inland.

Toronto is also one of the most multicultural cities in the world with more than 140 languages and dialects are spoken in the city, and almost half the population Toronto were born outside Canada.

Although not the capital city of the country – that particular honour rests with Ottawa – Toronto is nonetheless the centre of many of Canada's industries, and therefore it offers many economic opportunities to new arrivals.

Expatriated executive to Toronto City looking for a new family home

Consistently ranked as one of the most liveable cities in the world, Toronto enjoys a reputation as an exciting, diverse, clean, and safe city to set up home. It has 50 kilometres of waterfront with beaches, parks, marinas and waterfront trail.

Selection criteria:

The key criteria to take into consideration his family needs and personal needs.

In order to simplify our quest:

we are going to consider as family needs the proximity of **good rated elementary schools** for his young children and the presence of plenty of **malls** for his wife.

With the intention of meeting his personal needs it would be the existence of **gyms** in the chosen neighbourhood

We are not going to consider the criminality factor due to Toronto is well known to have low crime rates for such a big city nor cost the cost of rental housing as it is included in the expatriation package benefits of the executive

Target Audience

What type of stakeholders would be interested in this project?

1. Investors who could benefit from the model to assess real estate investments in high qualified potential neighbourhoods
2. Commercial Real Estate Brokers (CBRE, Cushman & Wakefield, etc..) encouraged to offer commercial and brokerage services related to the new locations.

3. Big Corporations, with no presence in the city, but willing to expand their business and operate in the city. They would need to know the impact of key parameters to be taken into consideration in a relocation process for their expatriated candidates.

4. Public Administration who can grant immigration permits, get taxes from large groups and would like to consider the factors of attractiveness to right size their infrastructures

5. Toronto residents who could benefit from the assessment model to take data driven decisions

6. Individuals in expatriation situation who may have to face a similar situation

2. Data Preparation:

We'll install the necessary packages, if missing, as

beautifulsoup4 to scrape websites

geopy to geocode web services and to locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders and other data sources

folium to visualize data that's been manipulated in Python on an interactive leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map

and libraries as:

numpy # to handle data in a vectorized manner
pandas # for data analysis
json # to handle JSON files
Nominatim # to convert an address into latitude and longitude values
requests # to handle requests
json_normalize # to tranform JSON file into a pandas dataframe

matplotlib and associated # to plotting graphsmodules
sklearn # to use machine learning k-means at clustering stage
folium # to map rendering
geocoder # to get coordinates

For this project we need the following data:

- Toronto data that contains the list of Boroughs and Neighborhoods
    - Data source :

      https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
    - Description : This data set contains the required information. And we will use this data set to explore various neighbourhoods of Toronto city.

- Venues in each neighbourhood of Toronto city.
    - Data source : Fourquare API

      https://developer.foursquare.com/docs/resources/categories

    - Description : By using this API we will get all the venues in each neighborhood. We can filter these venues to get only those that meet the predefined criteria.

- GeoSpace data
    - Data source : *https://open.toronto.ca/dataset/community-council-boundaries/*
    - Description : By using this geo space data we will get the Toronto Borough boundaries.

To simplify this project, we will only use Toronto neighbourhoods where Borough contains Toronto.

In order to obtain ratings data for elementary schools we will use *ontario.compareschoolrankings.org*

2.1. Scrapping Toronto Neighbourhoods from Wikipedia

I first make use of Wikipedia on its page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M page

to scrap the table in order to create a data-frame.

For this, I've used requests and Beautifulsoup4library to create a data-frame containing the PostalCode, the Borough and the Neigbourhood

[29]:

	PostalCode	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Kingsway Park South West,Mimico NW,The Queensw...
3	M6A	North York	Lawrence Heights,Lawrence Manor
4	M7A	Queen's Park	Queen's Park
5	M9A	Etobicoke	Islington Avenue
6	M1B	Scarborough	Rouge,Malvern
7	M3B	North York	Don Mills North
8	M4B	East York	Woodbine Gardens,Parkview Hill
9	M5B	Downtown Toronto	Ryerson,Garden District
10	M6B	North York	Glencairn

2.2. Getting coordinates of Boroughs: Geopy Client

The following step is to get the coordinates of the Boroughs, and the 103 Neighbourhoods using geocoder class of Geopy client along with their latitude and longitude.

2.3. Using Foursquare Location Data:

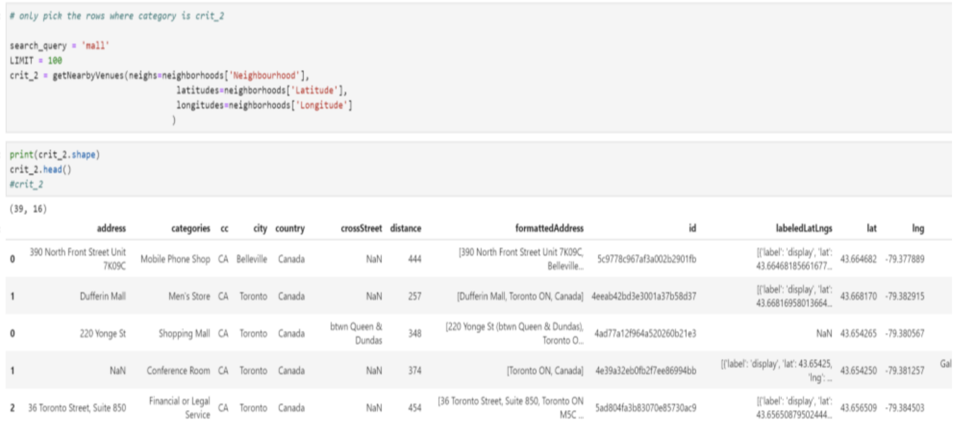
Foursquare data is very comprehensive and it powers location data for Apple, Uber etc.

Expatriated executive to Toronto City looking for a new family home

For this business problem I have used, as a part of the assignment, the Foursquare API to retrieve information about the pre-defined criteria to find the suitable Neighbourhood in Toronto

The call returns a JSON file that we need to turn that into a data-frame.

We repeated the process twice, once for the Gym criterium and another for the Mall.



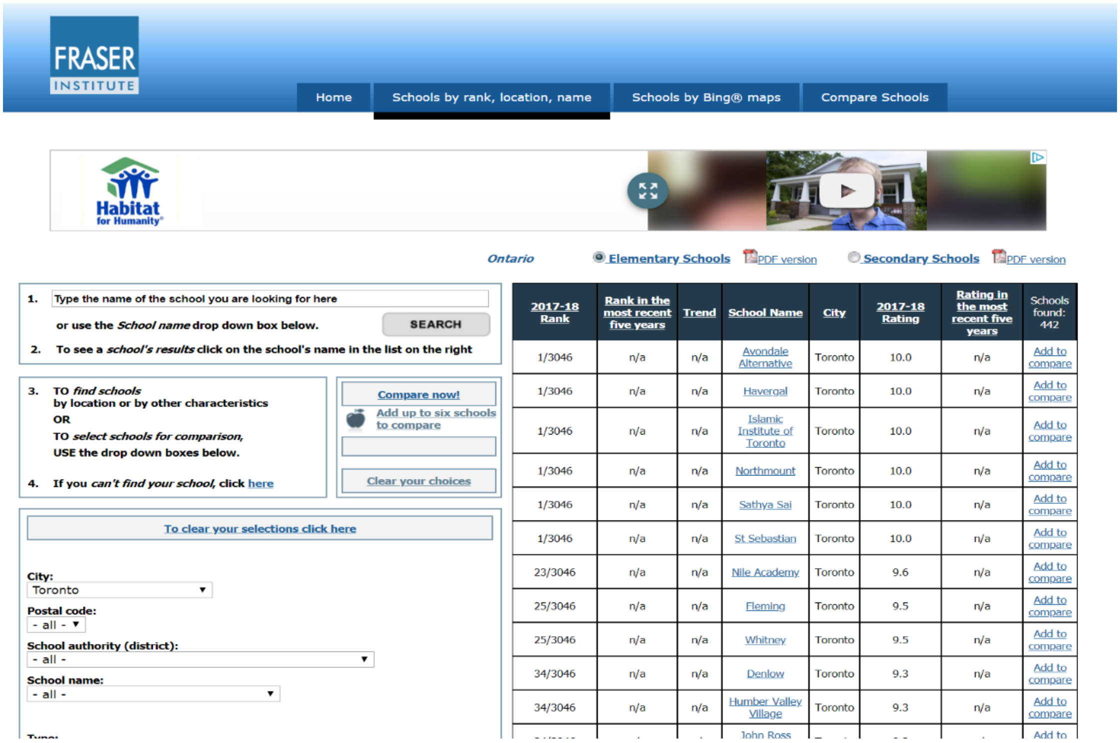
```
# only pick the rows where category is crit_2

search_query = 'mall'
LIMIT = 100
crit_2 = getNearbyVenues(neighs=neighborhoods['Neighbourhood'],
                         latitudes=neighborhoods['Latitude'],
                         longitudes=neighborhoods['Longitude']
                         )
```

```
print(crit_2.shape)
crit_2.head()
#crit_2
```

(39, 16)

	address	categories	cc	city	country	crossStreet	distance	formattedAddress	id	labeledLatLngs	lat	lng
0	390 North Front Street Unit 7K09C	Mobile Phone Shop	CA	Belleville	Canada	NaN	444	[390 North Front Street Unit 7K09C, Belleville...	5c9778c967af3a002b2901fb	[{'label': 'display', 'lat': 43.66468185661677...	43.664682	-79.377889
1	Dufferin Mall	Men's Store	CA	Toronto	Canada	NaN	257	[Dufferin Mall, Toronto ON, Canada]	4eeab42bd3e3001a37b58d37	[{'label': 'display', 'lat': 43.66816958013664...	43.668170	-79.382915
0	220 Yonge St	Shopping Mall	CA	Toronto	Canada	btwn Queen & Dundas	348	[220 Yonge St (btwn Queen & Dundas), Toronto O...	4ad77a12f964a520260b21e3	NaN	43.654265	-79.380567
1	NaN	Conference Room	CA	Toronto	Canada	NaN	374	[Toronto ON, Canada]	4e39a32eb0fb2f7ee86994bb	[{'label': 'display', 'lat': 43.65425, 'lng': ...	43.654250	-79.381257
2	36 Toronto Street, Suite 850	Financial or Legal Service	CA	Toronto	Canada	NaN	454	[36 Toronto Street, Suite 850, Toronto ON M5C ...	5ad804fa3b83070e85730ac9	[{'label': 'display', 'lat': 43.65650879502444...	43.656509	-79.384503

To get the ranking for the elementary schools we build a dataframe based on the scrapped information from



2017-18 Rank	Rank in the most recent five years	Trend	School Name	City	2017-18 Rating	Rating in the most recent five years	Schools found: 442
1/3046	n/a	n/a	Avondale Alternative	Toronto	10.0	n/a	Add to compare
1/3046	n/a	n/a	Havergal	Toronto	10.0	n/a	Add to compare
1/3046	n/a	n/a	Islamic Institute of Toronto	Toronto	10.0	n/a	Add to compare
1/3046	n/a	n/a	Northmount	Toronto	10.0	n/a	Add to compare
1/3046	n/a	n/a	Sathya Sai	Toronto	10.0	n/a	Add to compare
1/3046	n/a	n/a	St Sebastian	Toronto	10.0	n/a	Add to compare
23/3046	n/a	n/a	Nile Academy	Toronto	9.6	n/a	Add to compare
25/3046	n/a	n/a	Fleming	Toronto	9.5	n/a	Add to compare
25/3046	n/a	n/a	Whitney	Toronto	9.5	n/a	Add to compare
34/3046	n/a	n/a	Denlow	Toronto	9.3	n/a	Add to compare
34/3046	n/a	n/a	Humber Valley Village	Toronto	9.3	n/a	Add to compare

Until we get a dataframe with the name of the school, the postal code, the Borough and the Latest Rank position

```
top_schools = school_ratings.loc[school_ratings['Rating'] >= 8 ].re
top_schools.head()
```

	School_name	Rating	PostalCode	Borough	Latest_ Rank
0	Withrow Avenue	9.0	M4K	East Toronto	58/3046
1	Pape Avenue	8.0	M4K	East Toronto	284/3046
2	Georges-Étienne-Cartier	8.0	M4L	East Toronto	284/3046
3	Blythwood	9.1	M4N	Central Toronto	47/3046
4	Bedford Park	8.6	M4N	Central Toronto	123/3046

The above school information will be merged with the Foursquare information into a new dataframe to start the analysis and the K-means clustering.

3. Visualization and Data Exploration:

In the process of obtaining the **PostalCode, the Borough and the Neigbourhood** we have had to initiate a data wrangling process

From the data source

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

We get the raw data that we have to parse using *BeautifulSoup* library and saved as a panda data frame but we still have to refine the data

We apply the following rules:

Rule 1: Only process the cells that have an assigned borough. Ignore cells with a borough that is Not assigned

We have 77 records with no data that we decide to eliminate

```
PostalCode      77
Borough         77
Neighbourhood   77
dtype: int64
```

Rule 2 :If a cell has a borough but a Not assigned neighbourhood, then the neighbourhood will be the same as the borough so

```
# replacing 'Not assigned' neighborhoods with the name of the Borough
df1.loc[df1['Neighbourhood'] == 'Not assigned', 'Neighbourhood'] = df1['Borough']
```

Rule 3: More than one neighbourhood can exist in one postal code area, ... rows will be combined into one row with the neighbourhoods separated with a comma

```
postalcodes = df1['PostalCode'].nunique()
boroughs = df1['Borough'].nunique()
neighbourhoods= df1['Neighbourhood'].nunique()

Unique Postalcodes : 103
Unique Boroughs  : 11
Unique Neighbourhoods  :209
```

Once we have the PostalCode, the Borough and the Neigbourhood we need to add their coordinates from the source *http://cocl.us/Geospatial_data*

We merge the coordinate data and the Postal code data

```
### We plan to merge  df2 and df-pcodes but we need to rename the column "Postal Code" to "PostalCode" in order to do a proper merger
```

```
df_pcodes.columns = ['PostalCode', 'Latitude', 'Longitude']
df2=df2.sort_values('PostalCode')
df_pcodes.head()
```

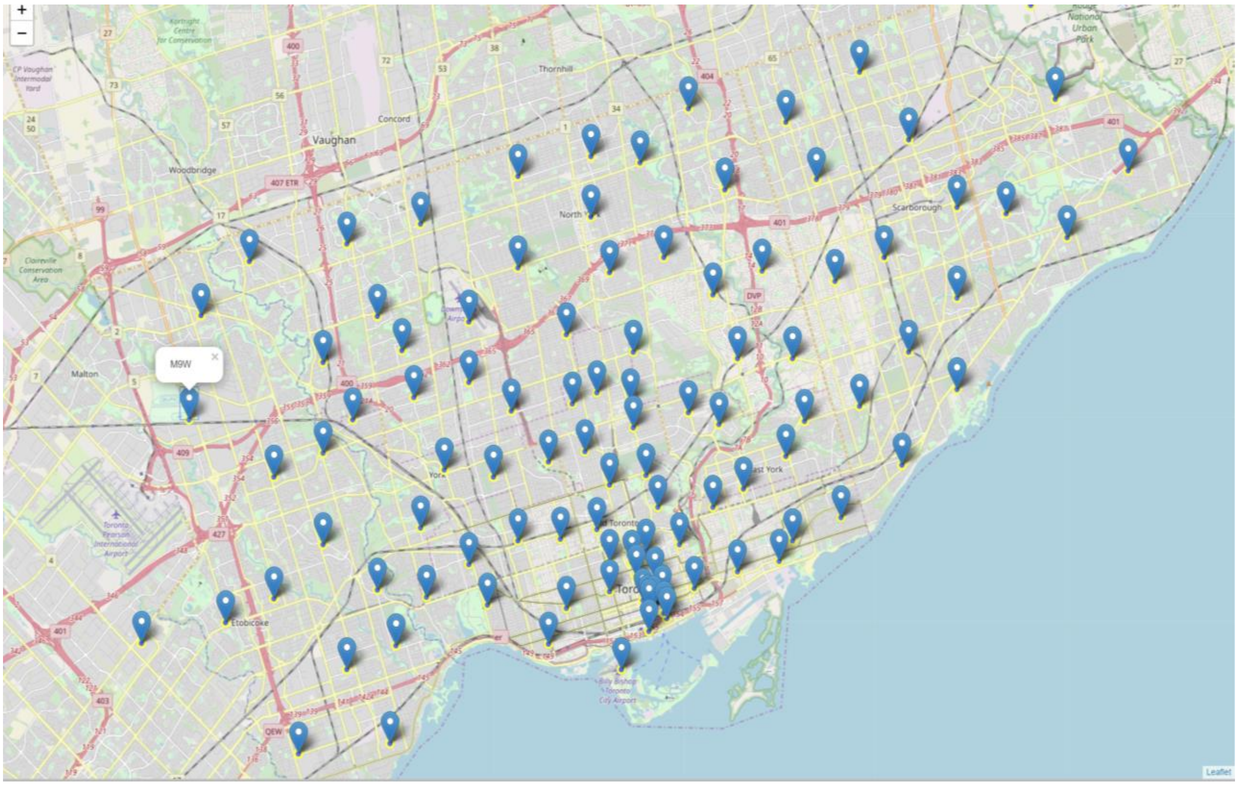
	PostalCode	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

```
# now we merge both files using PostalCode as common key
df3=pd.merge(df2,df_pcodes, how='left', on = 'PostalCode')
df3
```

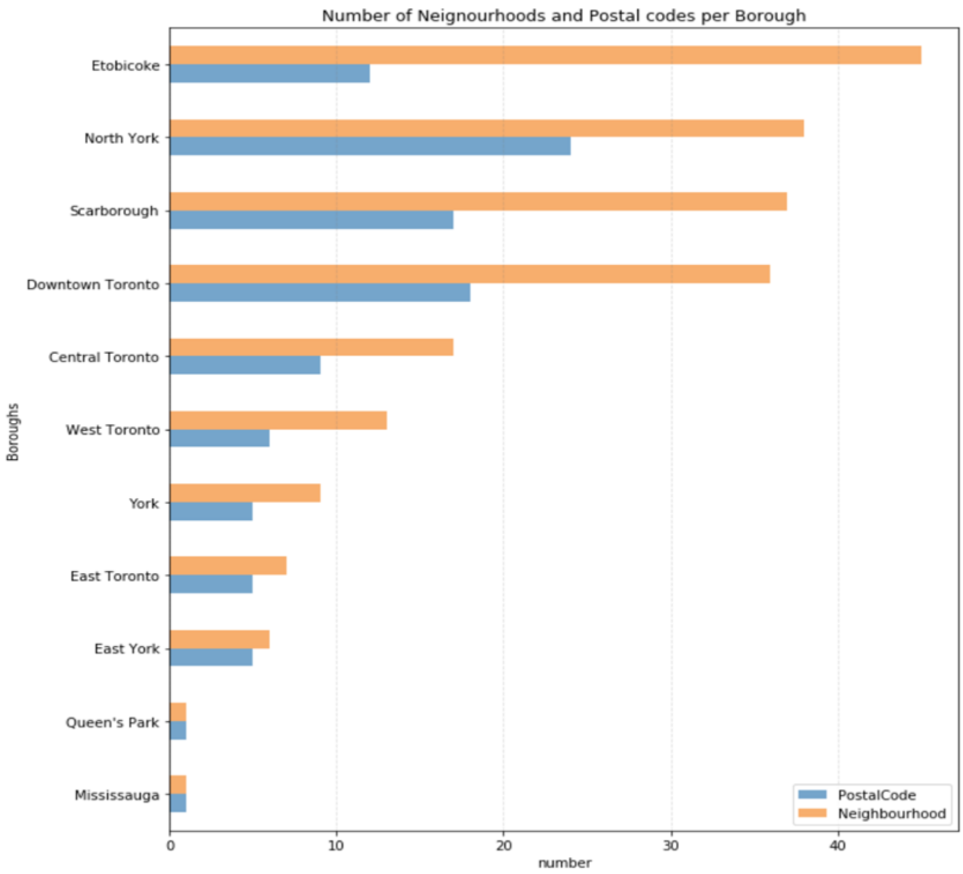
	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476

3.1. Exploratory Data Analysis:

Our starting point is a dataframe 103 rows x 5 columns, still quite "populated" in terms of postal codes



With the following distribution



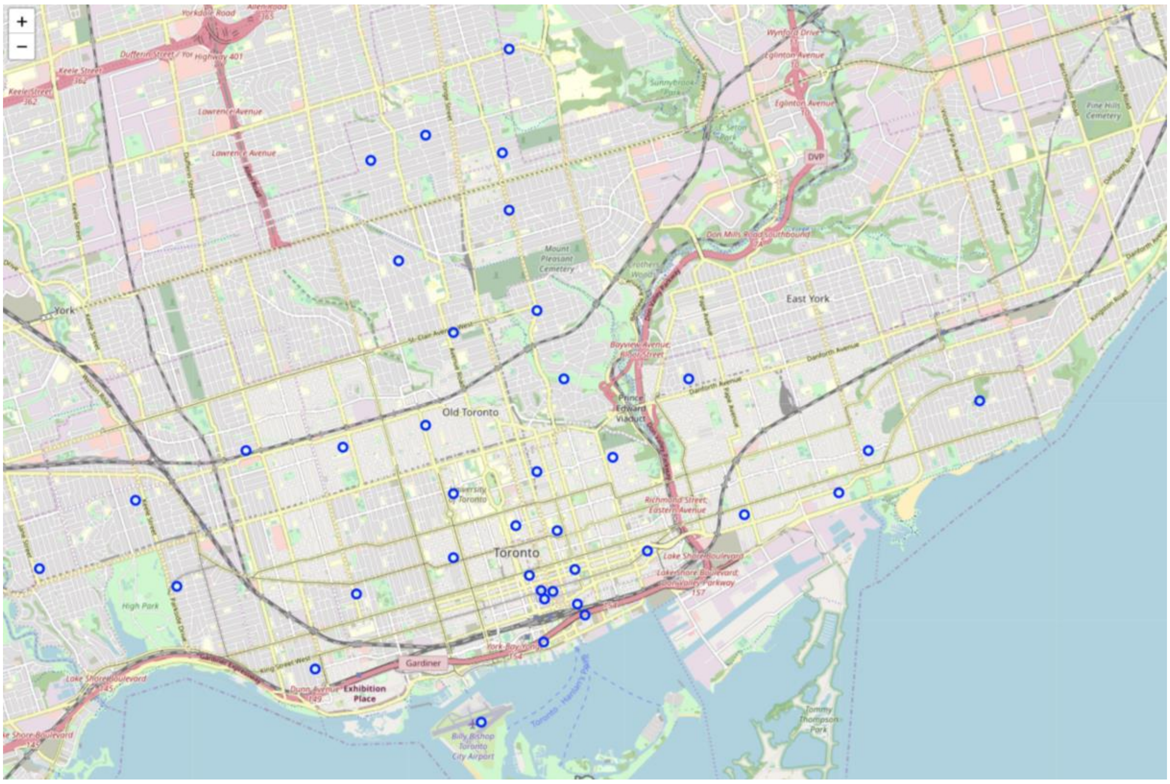
We're going to focus just on the " *** Toronto" area

```
From all Postal codes we will only get the ones that contain Toronto in the Borough.
```

```
neighborhoods = neighborhoods[neighborhoods['Borough'].str.contains("Toronto")]
neighborhoods
```

	PostalCode	Borough	Neighbourhood	Latitude	Longitude
37	M4E	East Toronto	The Beaches	43.676357	-79.293031
41	M4K	East Toronto	The Danforth West,Riverdale	43.679557	-79.352188
42	M4L	East Toronto	The Beaches West,India Bazaar	43.668999	-79.315572
43	M4M	East Toronto	Studio District	43.659526	-79.340923
44	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790
45	M4P	Central Toronto	Davisville North	43.712751	-79.390197
46	M4R	Central Toronto	North Toronto West	43.715383	-79.405678

That means a reduced dataset of 38 rows x 5 columns



The following step it is to start analysing each neighbourhood for the required relocation criteria

We'll use 2 built functions

One for the category type

```python
# function that extracts the category of the venue

def get_category_type(row):
    try:
        categories_list = row['categories']
    except:
        categories_list = row['venue.categories']

    if len(categories_list) == 0:
        return None
    else:
        return categories_list[0]['name']
```

And another for the nearby venues

```
# function that find nearby venues for a neighborhood based on a search query
def getNearbyVenues(neighs, latitudes, longitudes, radius=500):

    dataframe_filtered = pd.DataFrame()
    nearby_schools = pd.DataFrame()
    for neigh, lat, lng in zip(neighs, latitudes, longitudes):

        dataframe_filtered = dataframe_filtered[0:0]
        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/search?client_id={}&client_secret={}&ll={},{}&v={}&query={}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            lat,
            lng,
            VERSION,
            search_query,
            radius,
            LIMIT)


        # make the GET request
        results = requests.get(url).json()

        # assign relevant part of JSON to venues
        venues = results['response']['venues']
        if (venues == []): continue

        # tranform venues into a dataframe
        dataframe = json_normalize(venues)
        #dataframe.head()

        # keep only columns that include venue name, and anything that is associated with location
        filtered_columns = ['name', 'categories'] + [col for col in dataframe.columns if col.startswith('location.')] + ['id']
        dataframe_filtered = dataframe.loc[:, filtered_columns]

        # filter the category for each row
        dataframe_filtered['categories'] = dataframe_filtered.apply(get_category_type, axis=1)

        # clean column names by keeping only last term
        dataframe_filtered.columns = [column.split('.')[-1] for column in dataframe_filtered.columns]
        dataframe_filtered['neighborhood'] = neigh

        nearby_schools = nearby_schools.append(dataframe_filtered)

    return(nearby_schools)
```

The criteria used in Foursquare produced

1) For the GYM, we get 261 results

```
]: print(crit_1.shape)
   crit_1.head()
```

```
(261, 16)
```

]:	address	categories	cc	city	country	crossStreet	distance	formattedAddress	id	labeledLatLngs	lat	lng	
0	1048 Broadview Ave.	Gym / Fitness Center	CA	Toronto	Canada	NaN	679	[1048 Broadview Ave., Toronto ON M4K 2B8, Canada]	539a6357498ef1b1c9b888e7	[{'label': 'display', 'lat': 43.684524, 'lng':...	43.684524	-79.357102	
0	Carlaw Ave	Gym / Fitness Center	CA	Toronto	Canada	Carlaw & Dundas	432	[Carlaw Ave (Carlaw & Dundas), Toronto Ont, Ca...	52c9880b498edccef881e7b0	[{'label': 'display', 'lat': 43.6634116374049,...	43.663412	-79.341104	
1	233 Carlaw Ave.	Gym	CA	Toronto	Canada	btwn Dundas St & Queen St. E	382	[233 Carlaw Ave. (btwn Dundas St & Queen St. E...	4cdf14f2f8a4a1434ae3dbbc	[{'label': 'display', 'lat': 43.66293178257155...	43.662932	-79.340321	
0	140 Erskine	Gym	CA	Toronto	Canada	NaN	271	[140 Erskine, Toronto ON, Canada]	4da99b34a86e771ea70e84c1	[{'label': 'display', 'lat': 43.71312601210131...	43.713126	-79.393537	
1	900 Mount Pleasant Road	Gym / Fitness Center	CA	Toronto	Canada	NaN	174	[900 Mount Pleasant Road, Toronto ON M4P 3J9, ...	4c3f2724db3b1b8d635e6695	[{'label': 'display', 'lat': 43.71167058860572...	43.711671	-79.391767	9

But we want to focus on wide sport activities so only pick the GYM/FITNESS CENTER subcategory reducing the results to 55

```
# only pick the rows where category is crit_1

crit_1_list = crit_1.loc[crit_1['categories'] == 'Gym / Fitness Center']
crit_1_list = crit_1_list[['id','name', 'categories','neighborhood']]
crit_1_count = crit_1_list.groupby('neighborhood').count().reset_index()
```

```
print(crit_1_list.shape)
crit_1_count
```

```
(55, 4)
```

	neighborhood	id	name	categories
0	Adelaide,King,Richmond	4	4	4
1	Berczy Park	1	1	1
2	Brockton,Exhibition Place,Parkdale Village	1	1	1
3	Central Bay Street	5	5	5
4	Church and Wellesley	11	11	11
5	Commerce Court,Victoria Hotel	2	2	2
6	Davisville North	2	2	2
7	Design Exchange,Toronto Dominion Centre	5	5	5
8	Dovercourt Village,Dufferin	2	2	2
9	First Canadian Place,Underground city	4	4	4
10	Forest Hill North,Forest Hill West	1	1	1
11	Harbord,University of Toronto	1	1	1
12	Harbourfront East,Toronto Islands,Union Station	2	2	2
13	Ryerson,Garden District	7	7	7
14	St. James Town	3	3	3
15	Stn A PO Boxes 25 The Esplanade	1	1	1
16	Studio District	1	1	1
17	The Annex,North Midtown,Yorkville	1	1	1
18	The Danforth West,Riverdale	1	1	1

For the MALL we get 39 results, in this case we do not further filter as we want diversity

```
print(crit_2.shape)
crit_2.head()
#crit_2
```

```
(39, 16)
```

	address	categories	cc	city	country	crossStreet	distance	formattedAddress	id	labeledLatLngs	lat	lng	name	neighborhood	postalCode	state
0	390 North Front Street Unit 7K09C	Mobile Phone Shop	CA	Belleville	Canada	NaN	444	[390 North Front Street Unit 7K09C, Belleville...	5c9778c967af3a002b2901fb	[{'label': 'display', 'lat': 43.66468185661677...	43.664682	-79.377889	Mobile Klinik Professional Smartphone Repair -...	Church and Wellesley	K8P 3C9	ON
1	Dufferin Mall	Men's Store	CA	Toronto	Canada	NaN	257	[Dufferin Mall, Toronto ON, Canada]	4eeab42bd3e3001a37b58d37	[{'label': 'display', 'lat': 43.66816958013664...	43.668170	-79.382915	Pinstripe	Church and Wellesley	NaN	ON
0	220 Yonge St	Shopping Mall	CA	Toronto	Canada	btwn Queen & Dundas	348	[220 Yonge St (btwn Queen & Dundas), Toronto O...	4ad77a12f964a520260b21e3	NaN	43.654265	-79.380567	CF Toronto Eaton Centre	Ryerson,Garden District	M5B 2H1	ON
1	NaN	Conference Room	CA	Toronto	Canada	NaN	374	[Toronto ON, Canada]	4e39a32eb0fb2f7ee86994bb	[{'label': 'display', 'lat': 43.65425, 'lng': ...	43.654250	-79.381257	Galleria Offices (above Eaton Centre Mall)	Ryerson,Garden District	NaN	ON
2	36 Toronto Street, Suite 850	Financial or Legal Service	CA	Toronto	Canada	NaN	454	[36 Toronto Street, Suite 850, Toronto ON M5C ...	5ad804fa3b83070e85730ac9	[{'label': 'display', 'lat': 43.65650879502444...	43.656509	-79.384503	Susan Mallin	Ryerson,Garden District	M5C 2C5	ON

```
crit_2_list = crit_2.loc[crit_2['categories'] != 'NaN']
crit_2_list = crit_2_list[['id','name', 'categories','neighborhood']]
crit_2_count = crit_2_list.groupby('neighborhood').count().reset_index()
```

```
print(crit_2_list.shape)
crit_2_count
```

```
(39, 4)
```

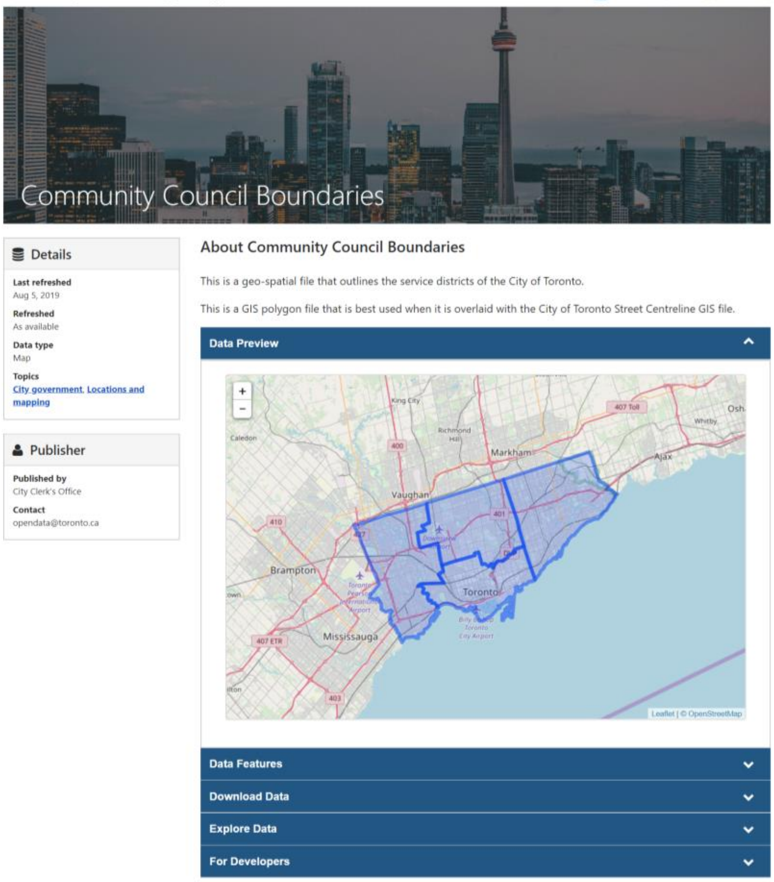
	neighborhood	id	name	categories
0	Adelaide,King,Richmond	3	3	3
1	Berczy Park	2	2	1
2	Brockton,Exhibition Place,Parkdale Village	1	1	1
3	Central Bay Street	3	3	3
4	Chinatown,Grange Park,Kensington Market	3	3	3
5	Church and Wellesley	2	2	2
6	Commerce Court,Victoria Hotel	3	3	2
7	Design Exchange,Toronto Dominion Centre	3	3	2
8	Dovercourt Village,Dufferin	7	7	7
9	First Canadian Place,Underground city	3	3	2
10	Harbourfront East,Toronto Islands,Union Station	1	1	1
11	High Park,The Junction South	1	1	1
12	Ryerson,Garden District	3	3	3
13	St. James Town	2	2	2
14	Stn A PO Boxes 25 The Esplanade	2	2	1

We analysed if we could the ratings in the above-mentioned categories but we reached the conclusion that it was not possible due to the lack of ratings

Due to many ratings for the venues chosen as criteria to be met we will not use the rating values from Foursquare data

```
def getVenueRating(venue_ids):

    ratingsl = []
    for venueid in zip(venue_ids):

        # create the API request URL
        venue_id = str(venueid)
        venue_id = venue_id.replace(',)','')
        venue_id = venue_id.replace('(','')
        venue_id = venue_id.strip('\'')
        #print(venue_id)

        url = 'https://api.foursquare.com/v2/venues/{}?client_id={}&client_secret={}&v={}'.format(
            venue_id,
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION)
        #print(url)
        # make the GET request
        result = requests.get(url).json()

        try:
            x = result['response']['venue']['rating']
        except:
            x = 0

        #print(x)
        ratingsl.append(x)

    return ratingsl
```

The key criteria in the neighbourhood selection is the school selection



In the Toronto open data website, we can find all the schools in the Toronto area but no information about its quality son we will use *ontario.compareschoolrankings.org* despite we have to build a dataset

Microsoft Excel csv Worksheet

3. Find best rated elementary schools in Toronto by neighborhood

As we want to use schools ratings we have to build a school dataset with Postal codes information using the raw data from 'ontario.compareschoolrankings.org'

```
# we upload the built data set to labs

school_ratings = pd.read_csv('EscuelasToronto.csv',sep=";")
school_ratings.head()
```

	School_name	Rating	PostalCode	Borough	Latest_Rank
0	Adam Beck	7.8	M4E	East Toronto	388/3046
1	Balmy Beach	7.6	M4E	East Toronto	495/3046
2	St Denis	7.2	M4E	East Toronto	740/3046
3	St John	7.0	M4E	East Toronto	865/3046
4	Williamson Road	6.7	M4E	East Toronto	1096/3046

But as we wanted good quality elementary schools, we picked only the highly rated

Only above 7

```
# As we want top quality schools we pick schools with hihg rating ( Above 7)

top_schools = school_ratings.loc[school_ratings['Rating'] >= 8 ].reset_index(drop=True)
top_schools.head()
```

	School_name	Rating	PostalCode	Borough	Latest_ Rank
0	Withrow Avenue	9.0	M4K	East Toronto	58/3046
1	Pape Avenue	8.0	M4K	East Toronto	284/3046
2	Georges-Étienne-Cartier	8.0	M4L	East Toronto	284/3046
3	Blythwood	9.1	M4N	Central Toronto	47/3046
4	Bedford Park	8.6	M4N	Central Toronto	123/3046

```
# We merge neighbourhoods data frame and tops schools dataframe on key Postalcode

top_schools_inToronto = top_schools.merge(neighborhoods, on='PostalCode', how = 'inner')
top_schools_inToronto
```

	School_name	Rating	PostalCode	Borough_x	Latest_ Rank	Borough_y	Neighbourhood	Latitude	Longitude
0	Withrow Avenue	9.0	M4K	East Toronto	58/3046	East Toronto	The Danforth West,Riverdale	43.679557	-79.352188
1	Pape Avenue	8.0	M4K	East Toronto	284/3046	East Toronto	The Danforth West,Riverdale	43.679557	-79.352188
2	Georges-Étienne-Cartier	8.0	M4L	East Toronto	284/3046	East Toronto	The Beaches West,India Bazaar	43.668999	-79.315572
3	Blythwood	9.1	M4N	Central Toronto	47/3046	Central Toronto	Lawrence Park	43.728020	-79.388790
4	Bedford Park	8.6	M4N	Central Toronto	123/3046	Central Toronto	Lawrence Park	43.728020	-79.388790
5	John Ross Robertson	9.3	M4R	Central Toronto	34/3046	Central Toronto	North Toronto West	43.715383	-79.405678
6	Davisville	8.6	M4S	Central Toronto	123/3046	Central Toronto	Davisville	43.704324	-79.388790
7	Whitney	9.5	M4T	Central Toronto	25/3046	Central Toronto	Moore Park,Summerhill East	43.689574	-79.383160

And finally, we can merge the dataset generated with Foursquare categories venue data, the postal code-Borough data and the schools ranking data

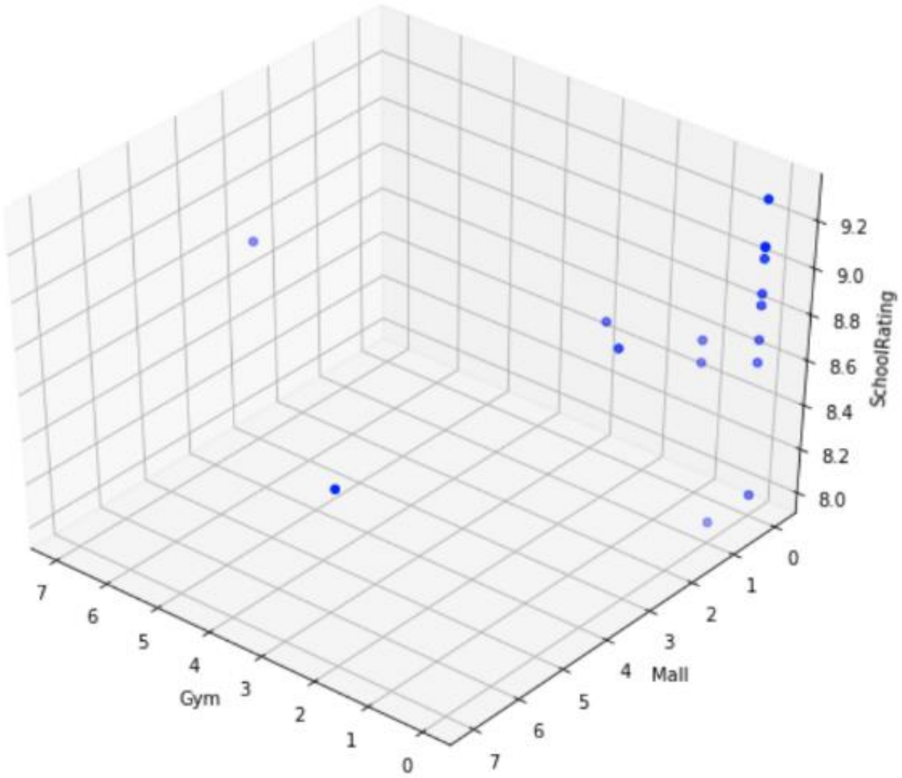
```
df = df[['Neighborhood','Latitude','Longitude','Rating','id_x','id_y']]
df.columns = ['Neighborhood','Latitude','Longitude','SchoolRating','GymCount','MallCount']
df.fillna(0, inplace = True)
df
```

	Neighborhood	Latitude	Longitude	SchoolRating	GymCount	MallCount
0	Berczy Park	43.644771	-79.373306	8.80	1.0	2.0
1	Christie	43.669542	-79.422564	8.00	0.0	0.0
2	Davisville	43.704324	-79.388790	8.60	0.0	0.0
3	Dovercourt Village,Dufferin	43.669005	-79.442259	8.70	2.0	7.0
4	Forest Hill North,Forest Hill West	43.696948	-79.411307	8.60	1.0	0.0
5	Harbourfront East,Toronto Islands,Union Station	43.640816	-79.381752	8.70	2.0	1.0

This will be our base information where to apply the machine learning algorithm, K- means clustering

As we have 3 variables, we can plot them in order to see how the distribution looks like

```
<mpl_toolkits.mplot3d.art3d.Path3DCollection at 0x7f6a3aeb7278>
```



4. Clustering the Boroughs

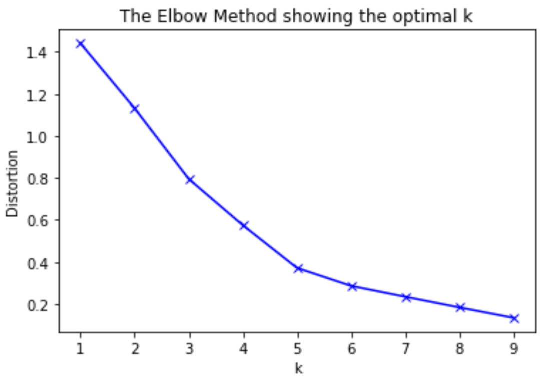
Now we can apply K-Means clustering but previously we need standardise our dataset values and to define what would be the appropiate number of clusters we have to generate

```
from sklearn.preprocessing import StandardScaler
F = df_g4.values[:,0:]
#F
```

```
F = np.nan_to_num(F)
G = StandardScaler().fit_transform(F)
#G
```

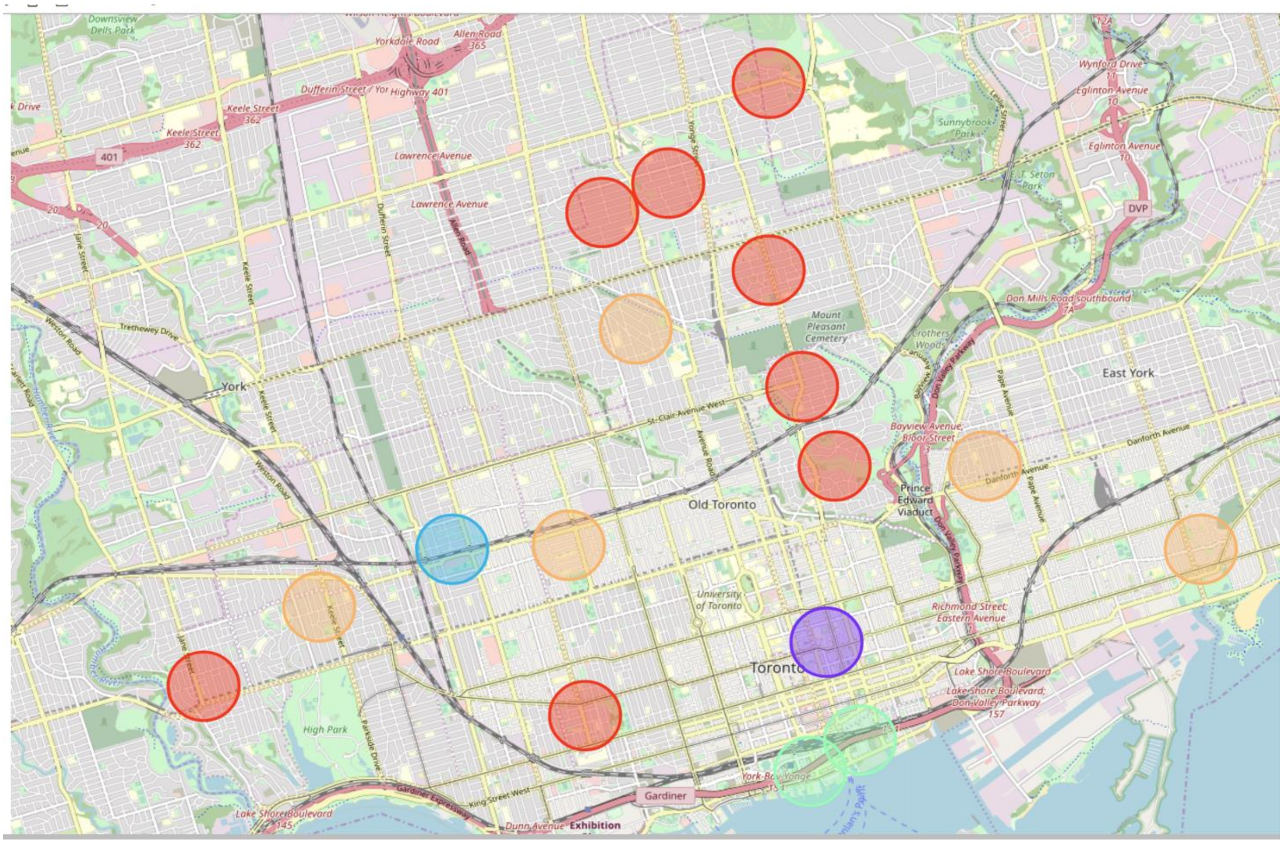
In order to determine the optimal k number, we'll apply **the elbow method**

```
# k means determine k
distortions = []
K = range(1,10)
for k in K:
    kmeanModel = KMeans(n_clusters=k).fit(G)
    kmeanModel.fit(G)
    distortions.append(sum(np.min(cdist(G, kmeanModel.cluster_centers_, 'euclidean'), axis=1)) / G.shape[0])
```



Finally, we try to cluster into 5 boroughs based on the 3 variables and use K-Means clustering. So, our expectation would be based on the similarities of venue categories, these districts will be clustered.

The 5 clusters of neighbourhoods spotted on the map



5. Results and Discussion:

We reached at the end of the analysis, where we got a sneak peak of the 5 major clusters of Toronto and, as the business problem started with benefits and drawbacks of choosing one new home for the family.

The data exploration was mostly concentrated on schools but also, I have used data from web resources like Wikipedia, python libraries like Geopy, and Foursquare API, to set up a very realistic data-analysis scenario.

We have found out that in Cluster 0 (red) we have the best ranked schools (above 8,6 and North Toronto West, Little Portugal, Trinity, Roselawn, Moore Park, and Summerhill East above 9, but no Gyms nor malls around so it could be a not very well balanced option and are clearly far away from downtown.

Cluster 4 (orange) it's the group that scores either on Gym either on Malls but never on both so discussions will be if father or mother wins.
Cluster3 (green) scores on both Gym and Mall, Harbourfront East, Toronto Islands, Union Station scores better on Gym and Berczy Park better on Malls, but they are close so it would be necessary to wander around the streets to refine the decision.

Cluster 1 (purple) and Cluster 2 (blue) offer above 8.7 in schools and higher scores in Gym and Mall than Cluster 3 but Cluster 1 scores more in Gym than Cluster 2 and vice versa so the final decision would require further information.

It's clear than many more criteria could be added to refine the decision adding more complexity and a pitfall of this analysis could be the consideration of only one major area of Toronto, taking into account all the areas may be would have provided a more realistic picture. Furthermore, this results also could potentially vary if we use some other clustering techniques like DBSCAN.

6. Conclusion

Finally, we have flavoured what could be a real-life data-science projects. We have applied some of the most frequently used python libraries to scrap web-data and used the Foursquare API to explore the neighbourhoods of Toronto and visualized the results of segmentation of the neighbourhoods using Folium leaflet map.

The exercise falls into the feasibility of this kind of analysis in real life business.

Also, some of the drawbacks and chance for improvements to represent even more realistic pictures are mentioned.

Toronto may be a good opportunity if someone is pursuing a a career in arts, culture, media or tech. It offers economic stability and opportunity in a variety of fields, with an increasing focus on tech., with offices for Google, Uber, Shopify, Vice magazine and more, there are over 200,000 tech and internet related jobs and counting.

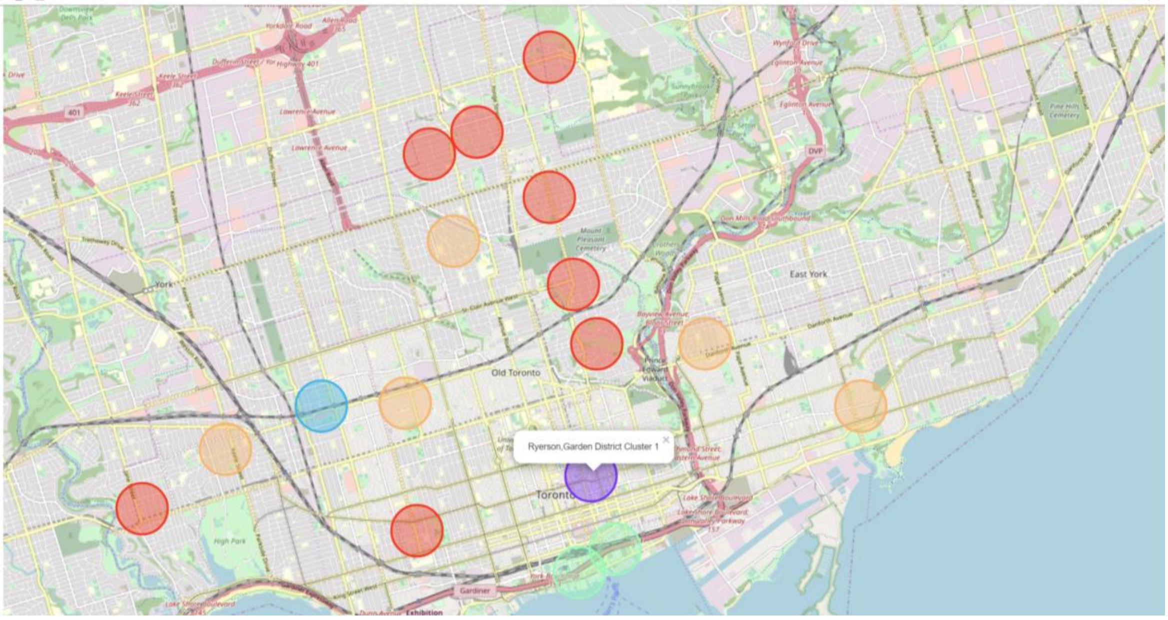
Toronto has a low crime rate, that's why we excluded this factor from the analysis.
In term of weather it's cold, a freeze compared to my home country, but unless you're coming from Vancouver or Victoria, you'd probably find Toronto's winters comparatively mild.
Housing is expensive, traffic and congestion are important issues, and long commute times as in every big city matter, but there are big parks around.

Families looking for a home big enough for 2 or 3 kids may find things difficult so that's why we use applied as school factor as a key matter in the exercise.

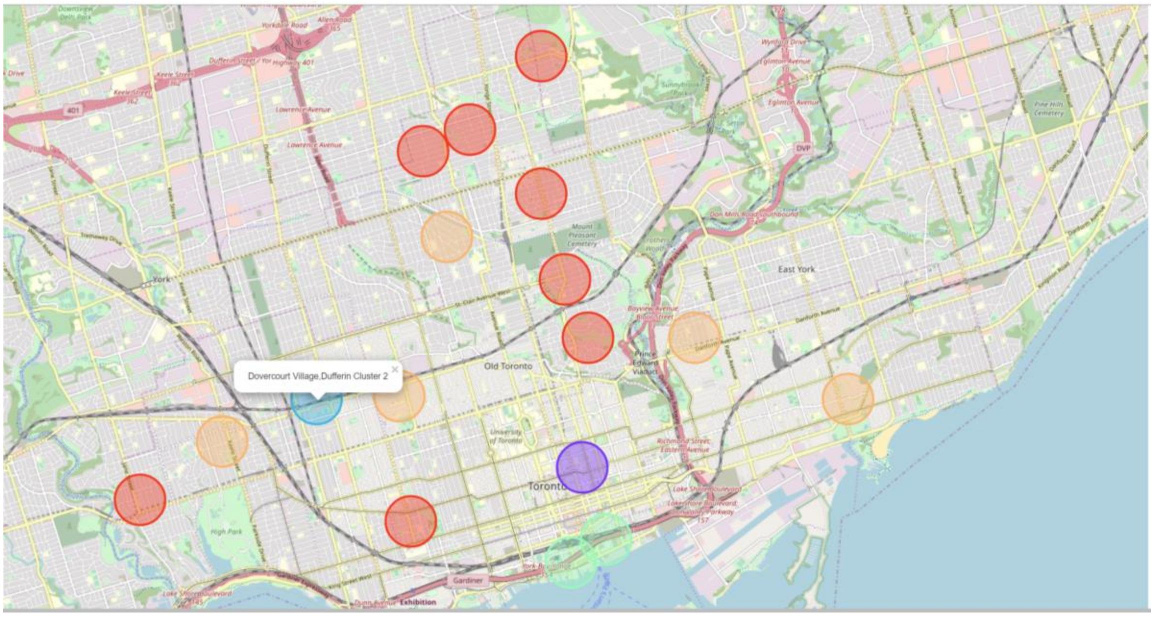
Expatriated executive to Toronto City looking for a new family home

According to the analysis the final options are

Cluster 1- Neighborhoods: Ryerson and Garden District



Cluster 2 -Neighborhoods: Dovercourt Village and Dufferin



A decision among them would require adding more criteria or stepping in those neighbourhoods to get a final decision.