Introduction

This project aims to explore and compare neighborhoods in Toronto for families who were forced to immigrate to Canada from different countries all over the world. Reasons for immigration are different, but choosing the right city to live in with a family and children is a difficult real-life problem.

Discussion of the background - problem description

Friends of mine - a couple with a two sons want to immigrate to Canada and someone advised them to settle down in Toronto.

The most important aspect for them is children. Their sons are 14 and 15 years old and should attend high school.

They also need to rent a house or apartment, which they can afford, and near a high school in a secure neighborhood.

The neighborhood also needs recreation places such as parks for everyday jogging and popular outdoor activities for families.

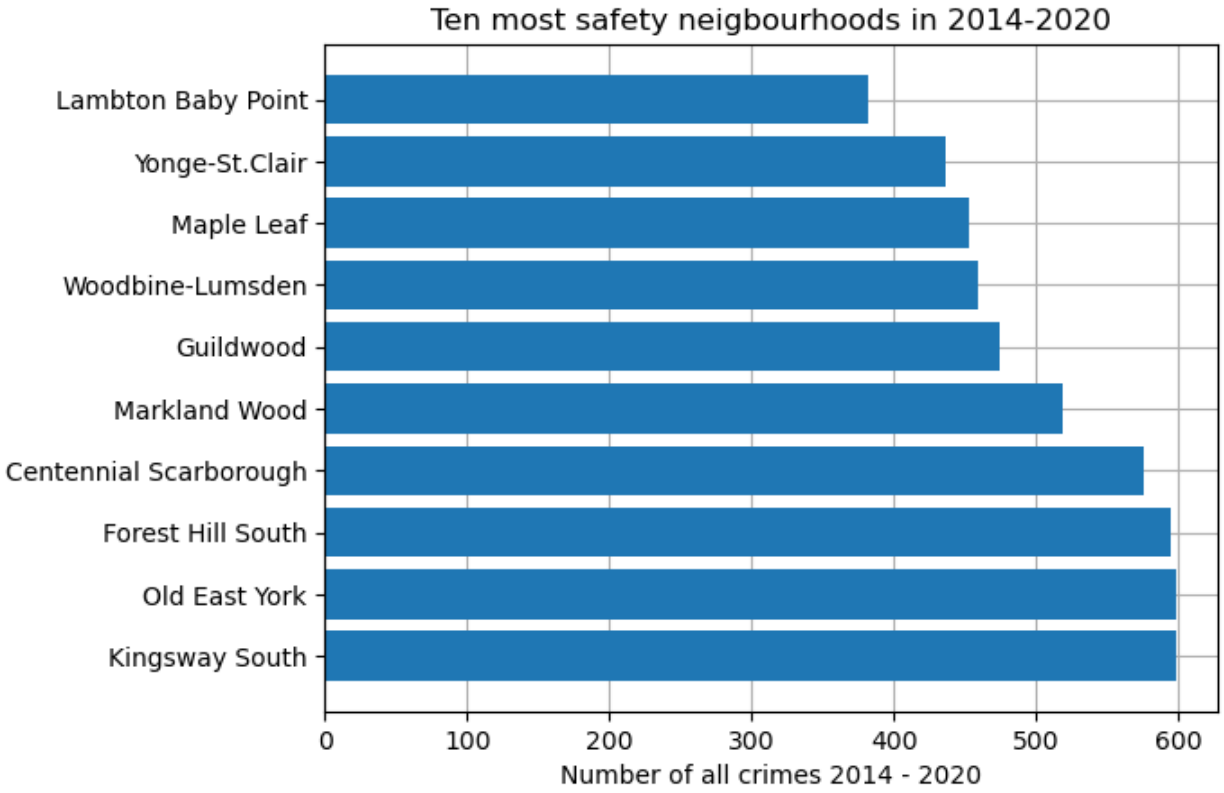
All analysis and a final report will be targeted at all families who are decided to move to Toronto and look for a nice place to live, not only for immigrants families.

Data section

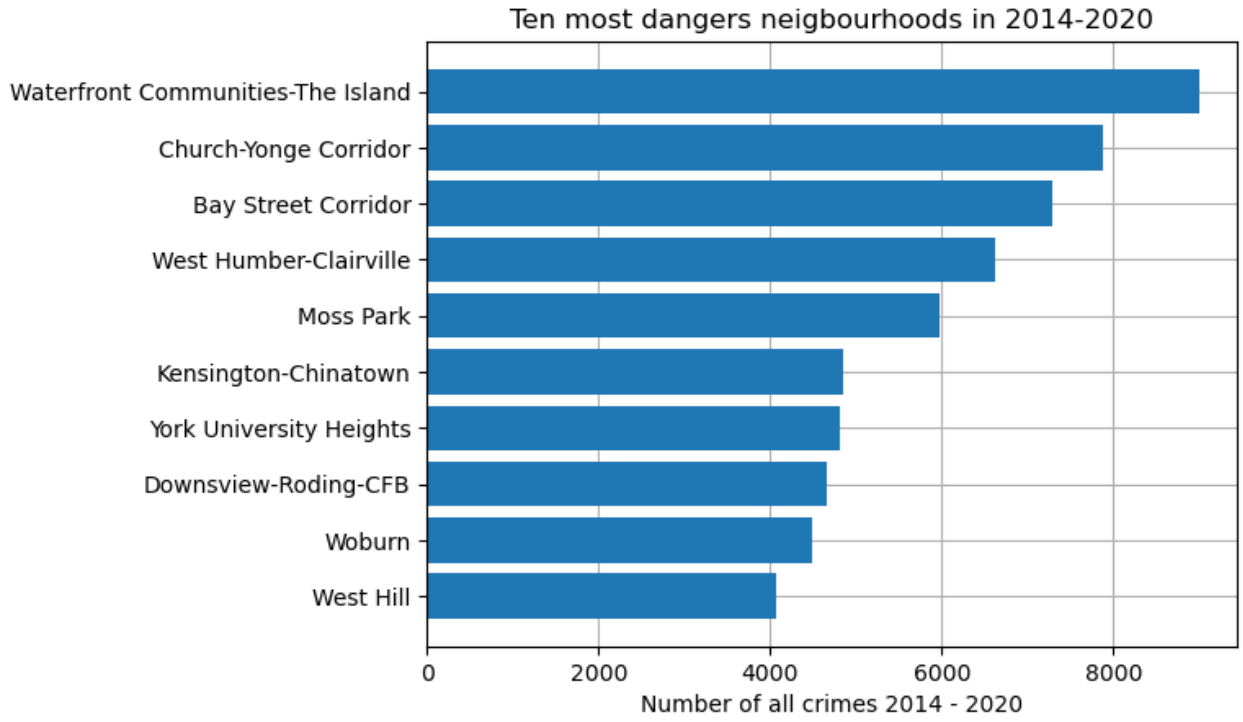
1. First, to get the list of neighborhoods in Toronto we use web scraping from Wikipedia page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
2. Using csv file to create the dataframe with the latitude and the longitude coordinates of each neighborhood GeoSpatial Dataset https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs_v1/Geospatial_Coordinates.csv
3. We noticed, that not all neighborhoods (140) are in including e.g. Yonge-St.Clair - and check with https://www.zipcodesonline.com/2020/06/postal-code-of-toronto-in-2020.html
4. Data about "Safety and crimes" we get from Toronto Police Service (Public Safety Data Portal) in *.csv file https://data.torontopolice.on.ca/datasets/3a1a9c98146e470e94e814b0e3a3fbca_0/explore?location=43.718404%2C-79.378190%2C11.73&showTable=true
5. Data about prices of houses in Toronto we get using web scraping average sale price of houses by neighborhoods in May 2021 from Zolo portal https://www.zolo.ca/toronto-real-estate/neighbourhoods and in 2020 from https://www.moneysense.ca/spend/real-estate/where-to-buy-real-estate-in-2020-city-of-toronto/. Because Zolo portal uses Cloudflare to secure website, we must download the webpage.
6. Information about secondary schools in Toronto we download data from Fraser Institute for secondary schools in Toronto https://www.tdsb.on.ca/Find-your/School/Secondary. Pages from Fraser Institute have only school names and scores - so we need to get school addresses. We extract some data from Toronto District School Board and Toronto Catholic District School Board.
7. We merging all data - number of crimes, house prices and secondary schools score, and Toronto neighborhoods - into one data frame.
8. We also use Foursquare API : "https://developer.foursquare.com/" to get all the venues in each neighborhood

Methodology section

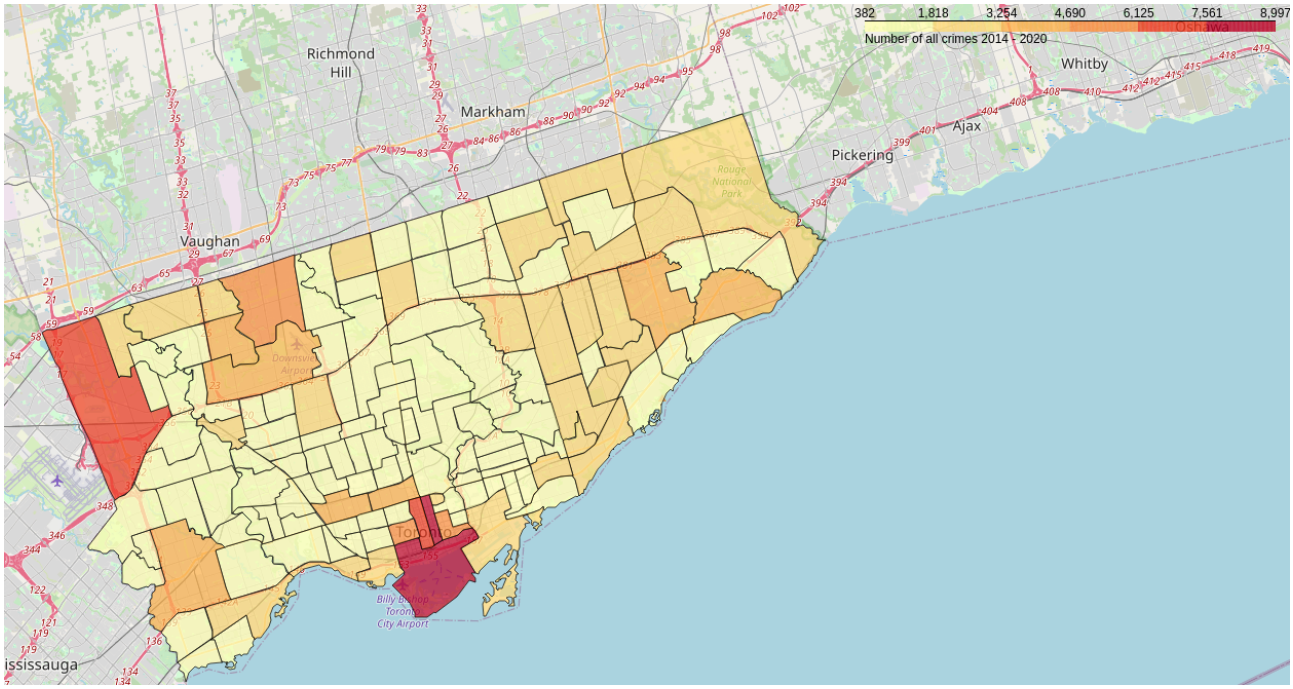
1. We start to analyze safety neighborhoods in Toronto. Using our data we calculated the sum of all crimes by neighborhoods in 2014-2020. We find ten of the safest neighborhoods:



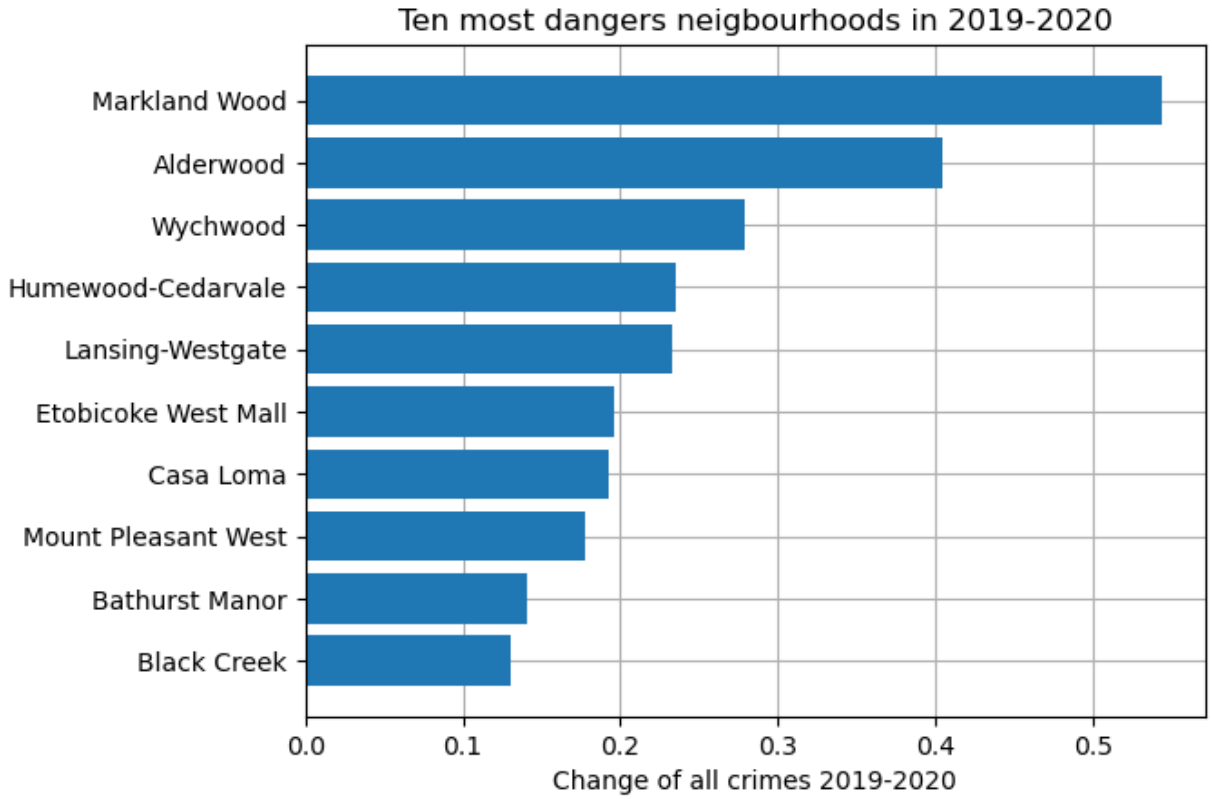
and ten of the most dangerous neighborhoods:



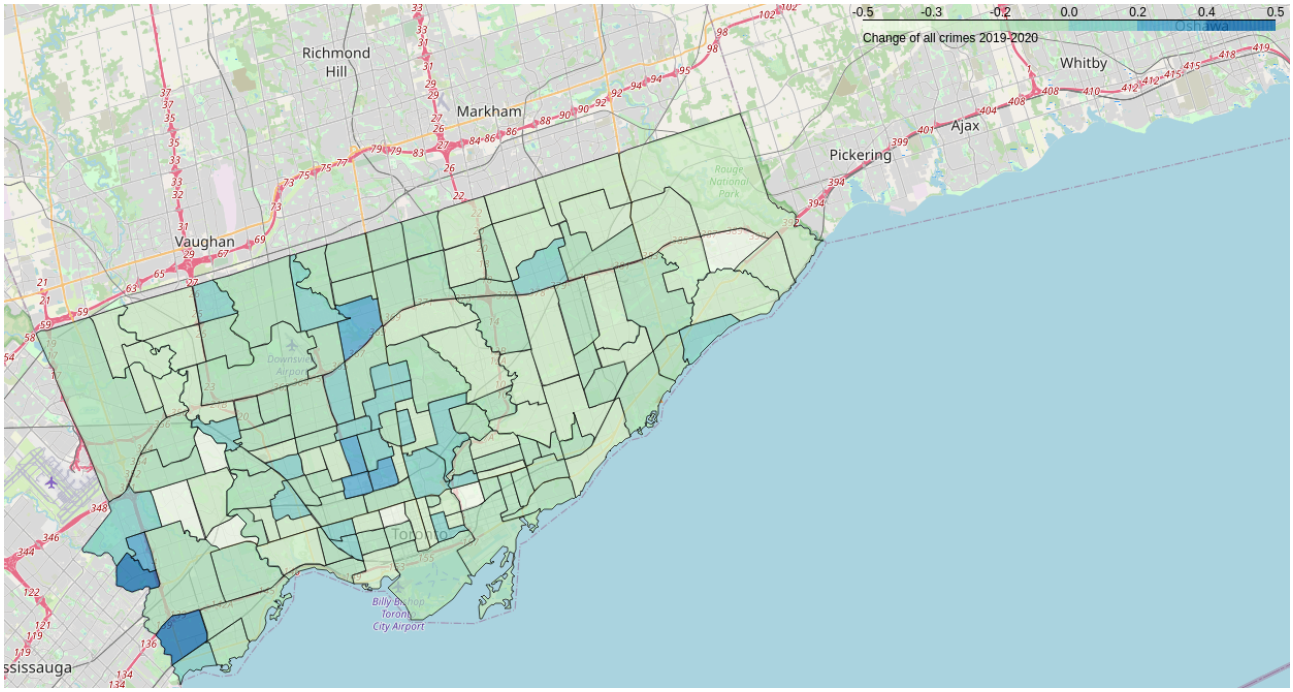
Then we presenting data using folium and choropleth library. To do this we use "Neighbourhood_Crime_Rates_2020.geojson" file from Toronto Police Service.



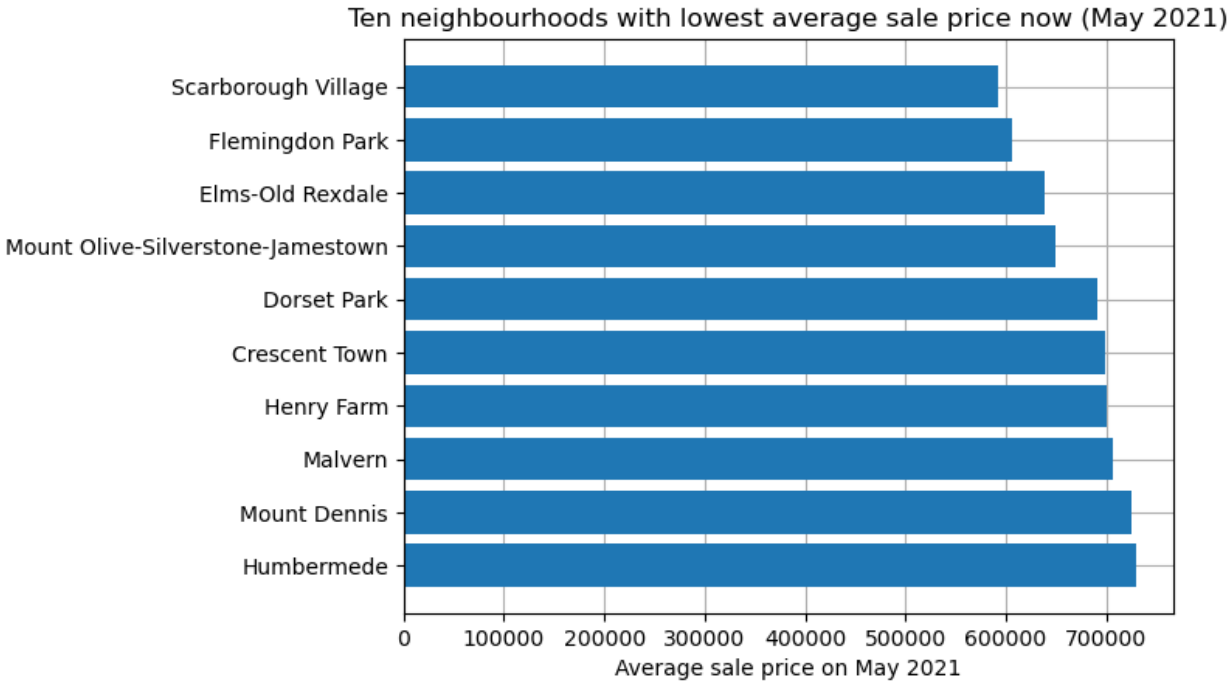
We also calculated the percent of change in crime numbers in 2019-2020 and ten of the most dangerous neighborhoods are:



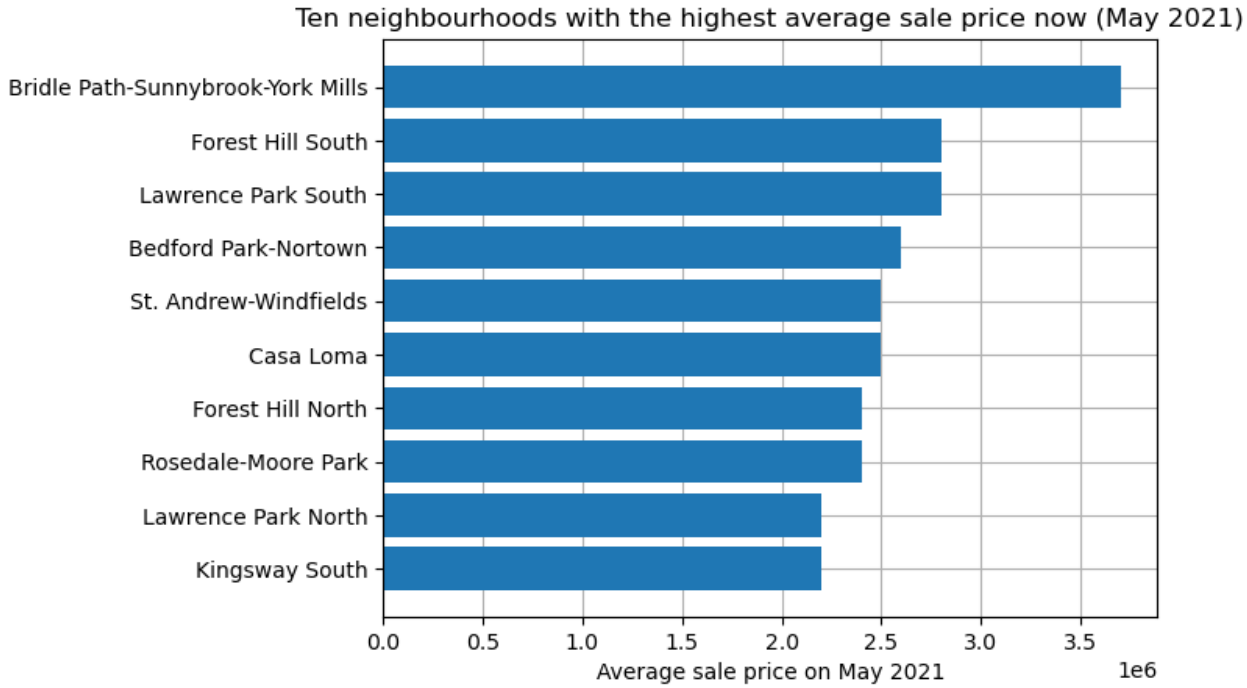
All tendencies we showed on the map.



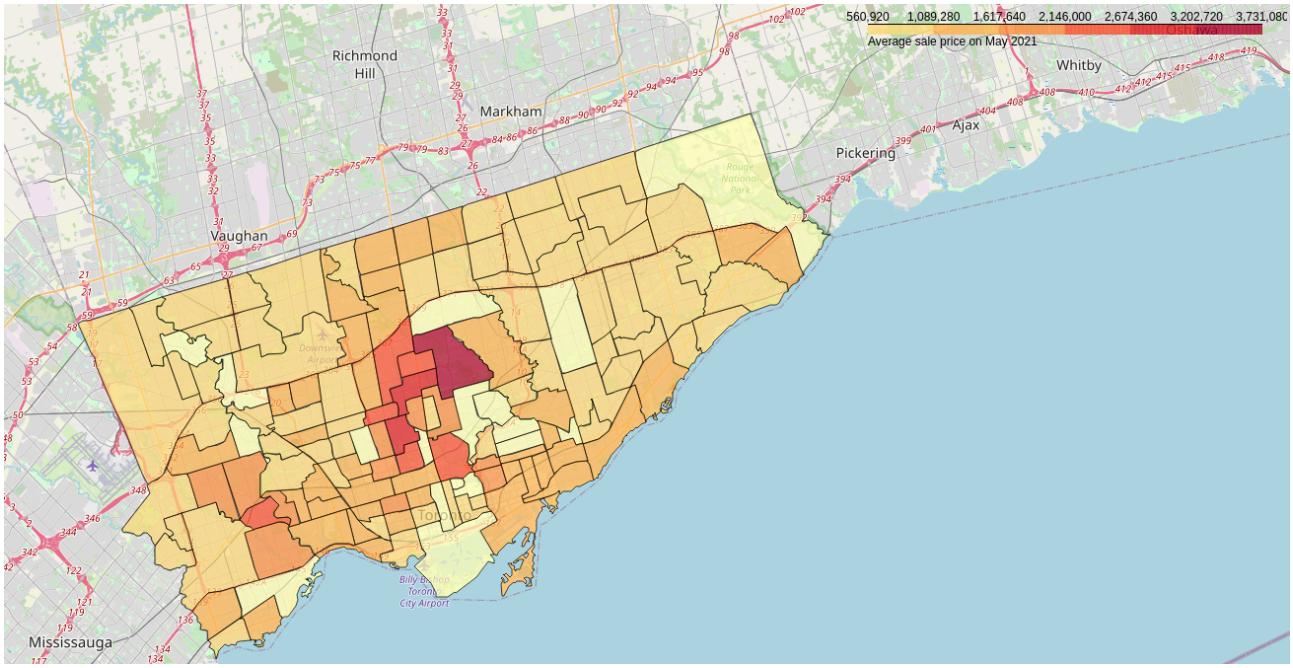
2. In the same manner we analyze house prices by neighborhoods in Toronto in 2020 - May 2021 with ten neighborhoods with the lowest average sale price now (May 2021):



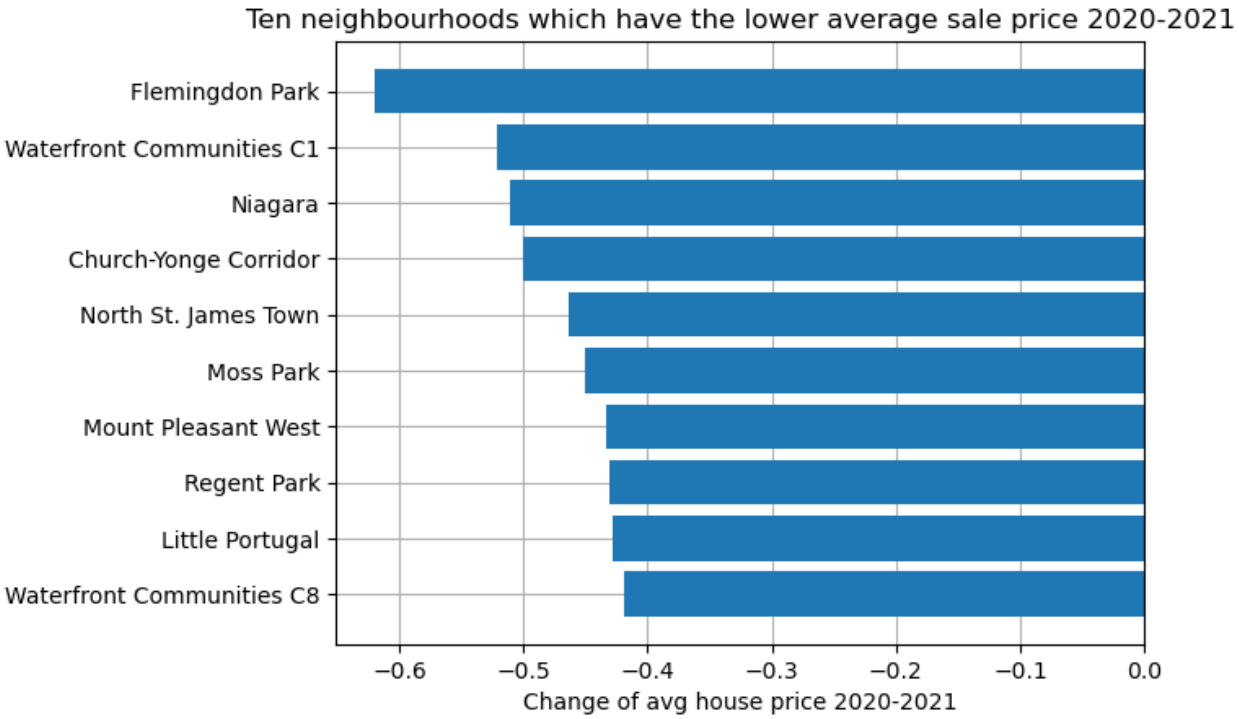
and the highest average sale price now (May 2021):



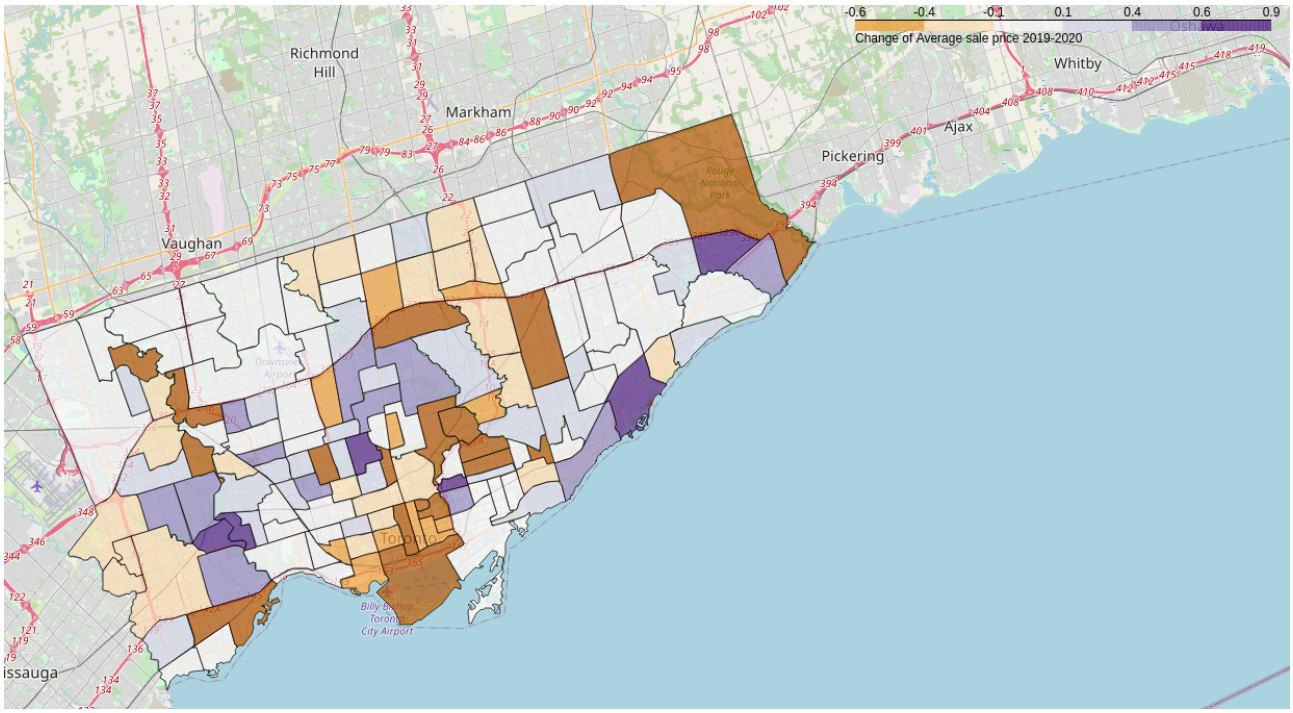
Then we presenting data using folium and choropleth library on a map.



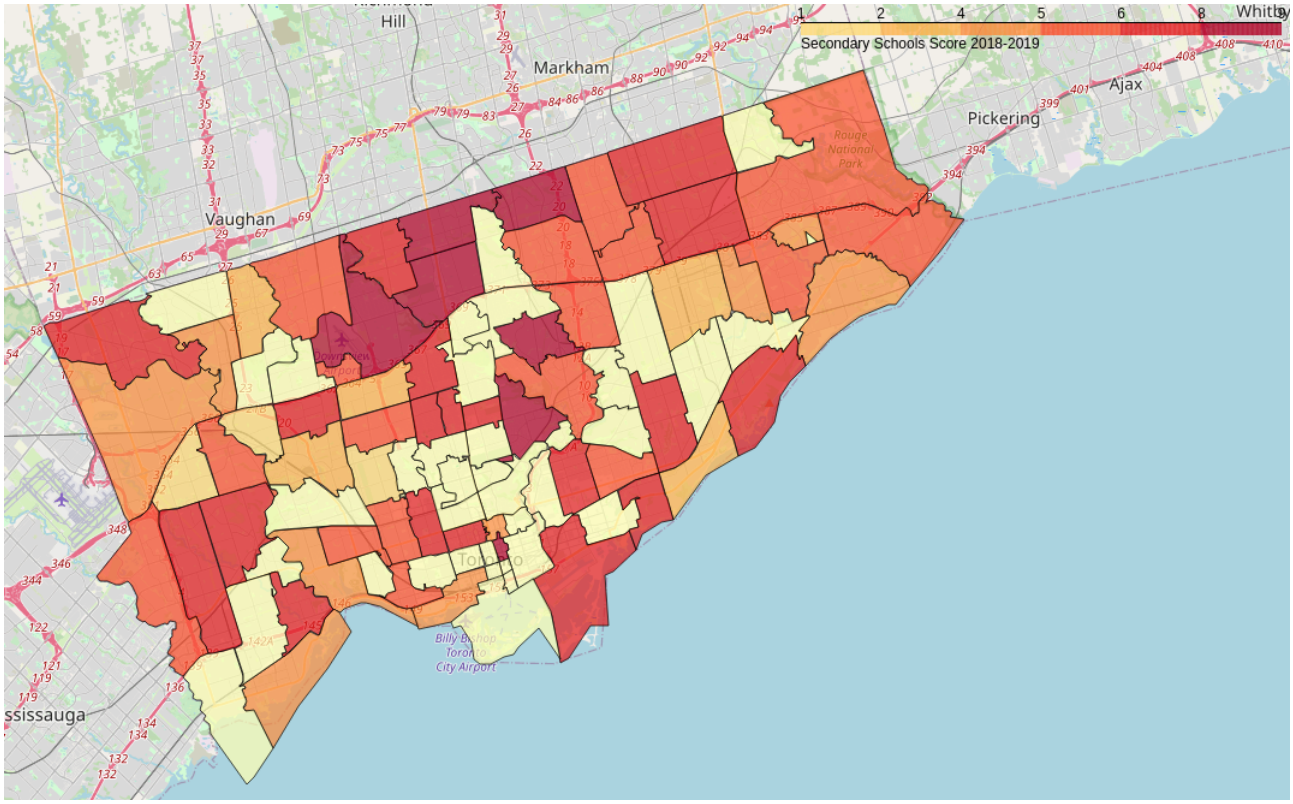
We also calculated the percent of change of average house price from 2020 to May 2021 with ten neighborhoods which in 2020-2021 have the lower average sale price:



Presenting data on a map:



3. We created a map of neighborhoods with secondary schools (according to a score)



4. We try to clustering using only three columns 'Number of all crimes 2014 - 2020','School Score 2018-2019','Average sale price - now'. Because the values were so different, we need to transform all data values to a set of real numbers from 0 to 1.

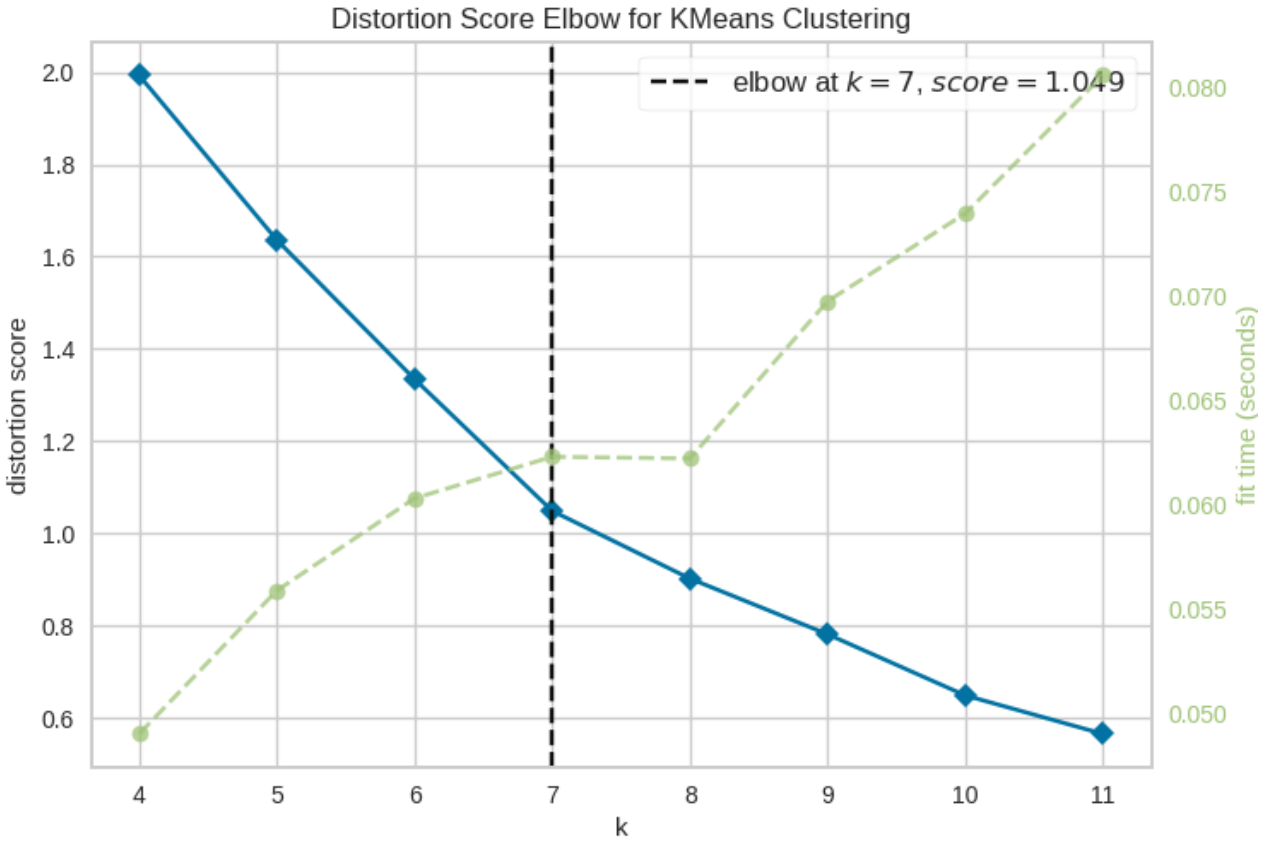
	Postalcode	Borough	Neighborhood	Latitude	Longitude	Number of all crimes 2014 - 2020	School Score 2018-2019	Average sale price - now
0	M3A	North York	Parkwoods,Parkwoods,Parkwoods-Donalda	43.753259	-79.329656	2014.0	6.0	1000000.0
1	M6A	North York	Lawrence Manor , Lawrence Heights,Briar Hill-B...	43.718518	-79.464763	1210.0	3.0	848000.0
2	M9A	Etobicoke	Islington Avenue,Edenbridge-Humber Valley,Humb...	43.667856	-79.532242	729.0	7.3	1800000.0
3	M1B	Scarborough	Malvern , Rouge,Agincourt South-Malvern West,M...	43.806686	-79.194353	3051.0	5.5	707000.0
4	M3B	North York	Don MillsNorth,Banbury-Don Mills,Don Mills	43.745906	-79.352188	1402.0	8.1	1500000.0
5	M5B	Downtown Toronto	Garden District,Church-Yonge Corridor,Garden D...	43.657162	-79.378937	7877.0	9.4	754000.0
6	M1C	Scarborough	Rouge Hill , Port Union , Highland Creek,Cente...	43.784535	-79.160497	2448.0	5.9	1300000.0
7	M3C	North York	Don MillsSouth,Flemingdon Park	43.725900	-79.340923	1375.0	5.9	606000.0
8	M4C	East York	Woodbine Heights,Danforth,Woodbine Heights,Woo...	43.695344	-79.318389	459.0	5.3	1100000.0

	Number of all crimes 2014 - 2020	School Score 2018-2019	Average sale price - now
0	0.217745	0.613636	0.203187
1	0.110474	0.272727	0.127490
2	0.046298	0.761364	0.601594
3	0.356104	0.556818	0.057271
4	0.136091	0.852273	0.452191

But some values e.g. prices of houses have a negative influence on our clustering - we are interested in lower prices, not higher so - we change them subtracting values from 1.

	Number of all crimes 2014 - 2020	School Score 2018-2019	Average sale price - now
0	0.782255	0.613636	0.796813
1	0.889526	0.272727	0.872510
2	0.953702	0.761364	0.398406
3	0.643896	0.556818	0.942729
4	0.863909	0.852273	0.547809
5	0.000000	1.000000	0.919323
6	0.724350	0.602273	0.647410
7	0.867512	0.602273	0.993028
8	0.989726	0.534091	0.747012
9	0.981855	0.556818	0.796813
10	0.865110	0.397727	0.886454

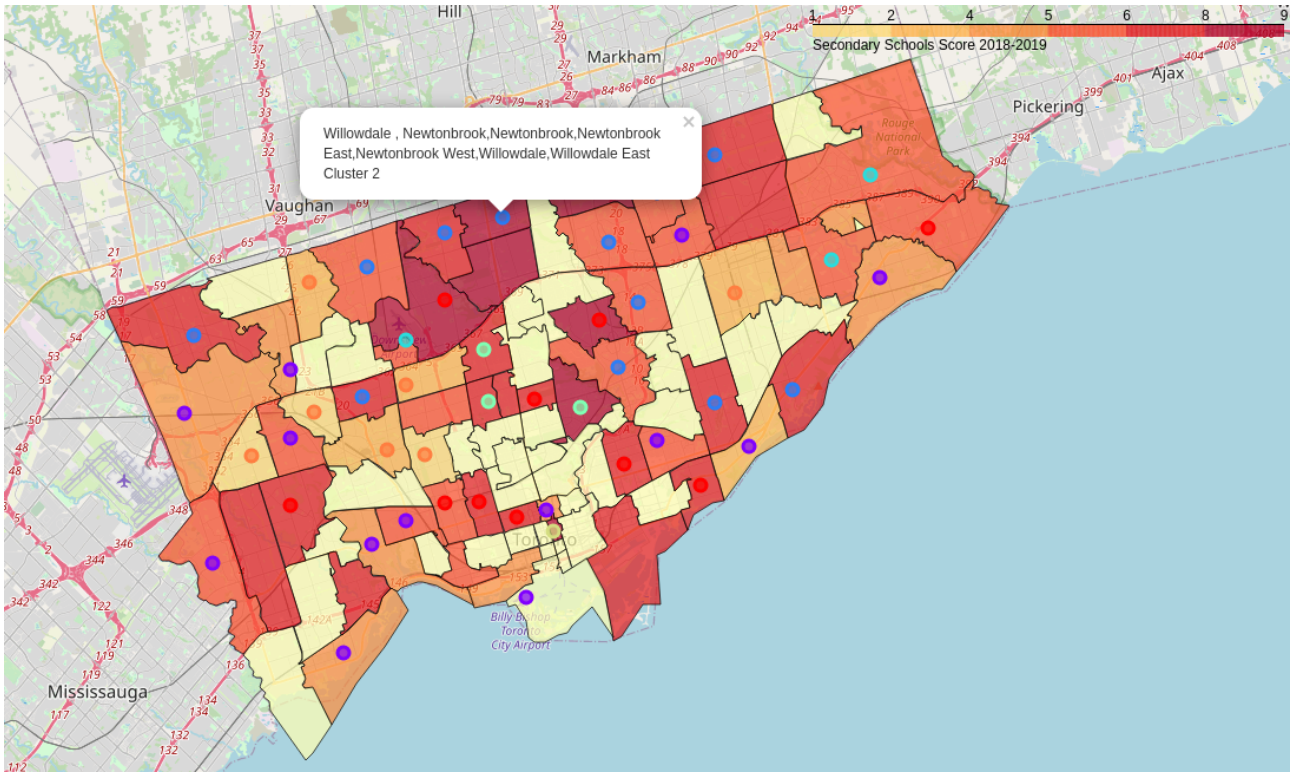
5. We use an unsupervised learning K-means algorithm to cluster the boroughs. First, we run K-Means to cluster the boroughs with k=5. Then we use Elbow Method using yellowbrick library to analyze the K-Means and get optimum k of the K-Means:



We get clusters:

	Postalcode	Borough	Neighborhood	Latitude	Longitude	Number of all crimes 2014 - 2020	School Score 2018-2019	Average sale price - now	Cluster Labels
0	M3A	North York	Parkwoods,Parkwoods-Donalda	43.753259	-79.329656	2014.0	6.0	1000000.0	2
1	M6A	North York	Lawrence Manor , Lawrence Heights,Briar Hill-B...	43.718518	-79.464763	1210.0	3.0	848000.0	6
2	M9A	Etobicoke	Islington Avenue,Edenbridge-Humber Valley,Humb...	43.667856	-79.532242	729.0	7.3	1800000.0	0
3	M1B	Scarborough	Malvern , Rouge,Agincourt South-Malvern West,M...	43.806686	-79.194353	3051.0	5.5	707000.0	3
4	M3B	North York	Don MillsNorth,Banbury-Don Mills,Don Mills	43.745906	-79.352188	1402.0	8.1	1500000.0	0

Visualization of the resulting clusters on a map with neighborhoods:



After examination, we find neighborhoods in Cluster 2 - as optimize neighborhoods in Toronto.

Cluster 2

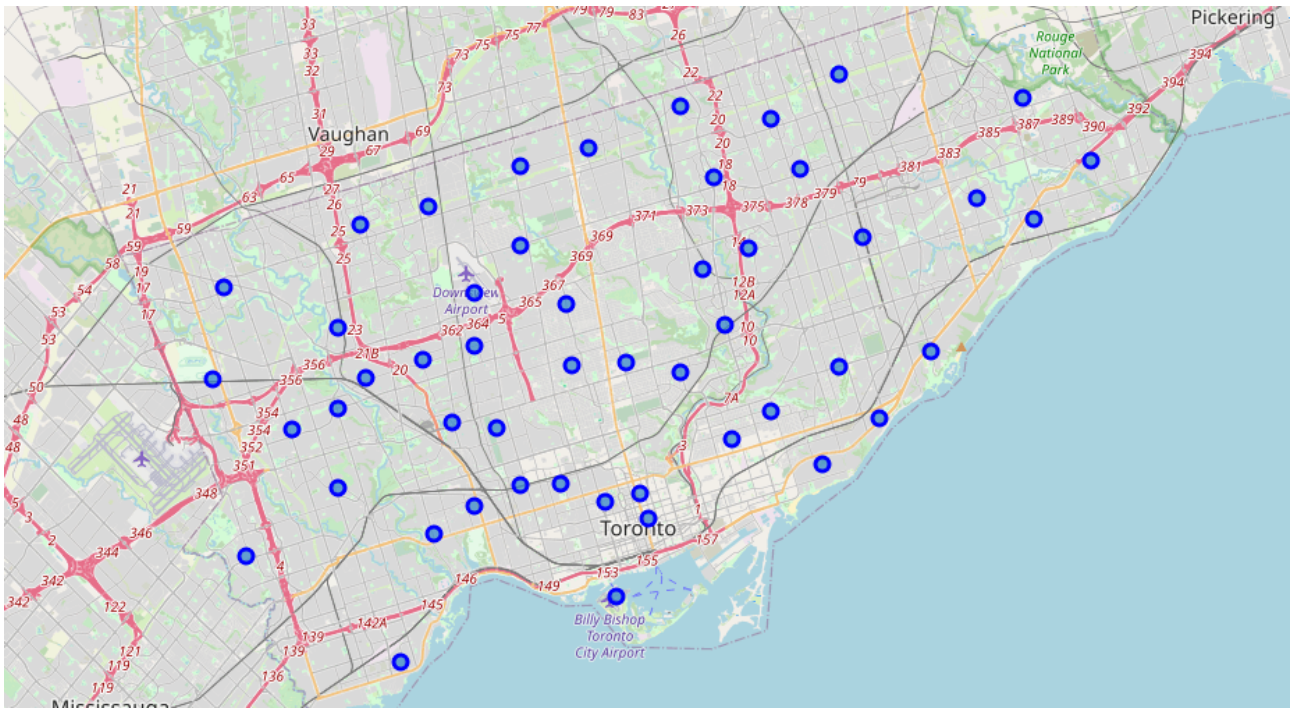
```
[95]: toronto_merged.loc[toronto_merged['Cluster Labels'] == 2, toronto_merged.columns[[1]+[2] + list(range(5, toronto_merged.shape[1]))]]
```

[95]:

	Borough	Neighborhood	Number of all crimes 2014 - 2020	School Score 2018-2019	Average sale price - now	Cluster Labels
0	North York	Parkwoods,Parkwoods,Parkwoods-Donalda	2014.0	6.0	1000000.0	2
7	North York	Don MillsSouth,Flemingdon Park	1375.0	5.9	606000.0	2
16	North York	Hillcrest Village,Bayview Woods-Steeles,Hillcr...	920.0	8.1	898000.0	2
19	North York	Fairview , Henry Farm , Oriole,Don Valley Vill...	703.0	6.3	1000000.0	2
20	North York	Northwood Park , York University,Northwood Par...	1764.0	6.0	805000.0	2
23	Scarborough	Golden Mile , Clairlea , Oakridge,Clairlea,Cla...	1958.0	7.8	892000.0	2
24	North York	North Park , Maple Leaf Park , Upwood Park,Map...	453.0	6.9	1200000.0	2
25	Scarborough	Cliffside , Cliffcrest , Scarborough Village W...	1459.0	7.1	592000.0	2
26	North York	Willowdale , Newtonbrook,Newtonbrook,Newtonbro...	1110.0	8.1	918000.0	2
38	North York	WillowdaleWest,Westminster-Branson,Willowdale ...	1098.0	7.2	1200000.0	2
43	Scarborough	Milliken , Agincourt North , Steeles East , L'...	2108.0	7.5	898000.0	2
46	Etobicoke	South Steeles , Silverstone , Humbergate , Jam...	920.0	6.9	898000.0	2
47	Scarborough	Steeles West , L'Amoreaux West,L'Amoreaux,L'Am...	920.0	6.2	898000.0	2

The most optimum place to settle down in Toronto would be neighborhoods from cluster 2. Secondary schools with high score 5.9 to 8.1, number of crimes 453 to 2108, avg house prices $ 592000 to $ 1200000

6. As the last thing we explore and clustering the neighborhoods in Toronto, besides high school score, number of crimes,avg house prices using venus which we get from the Foursquare API.
   We created a map with neighborhoods:



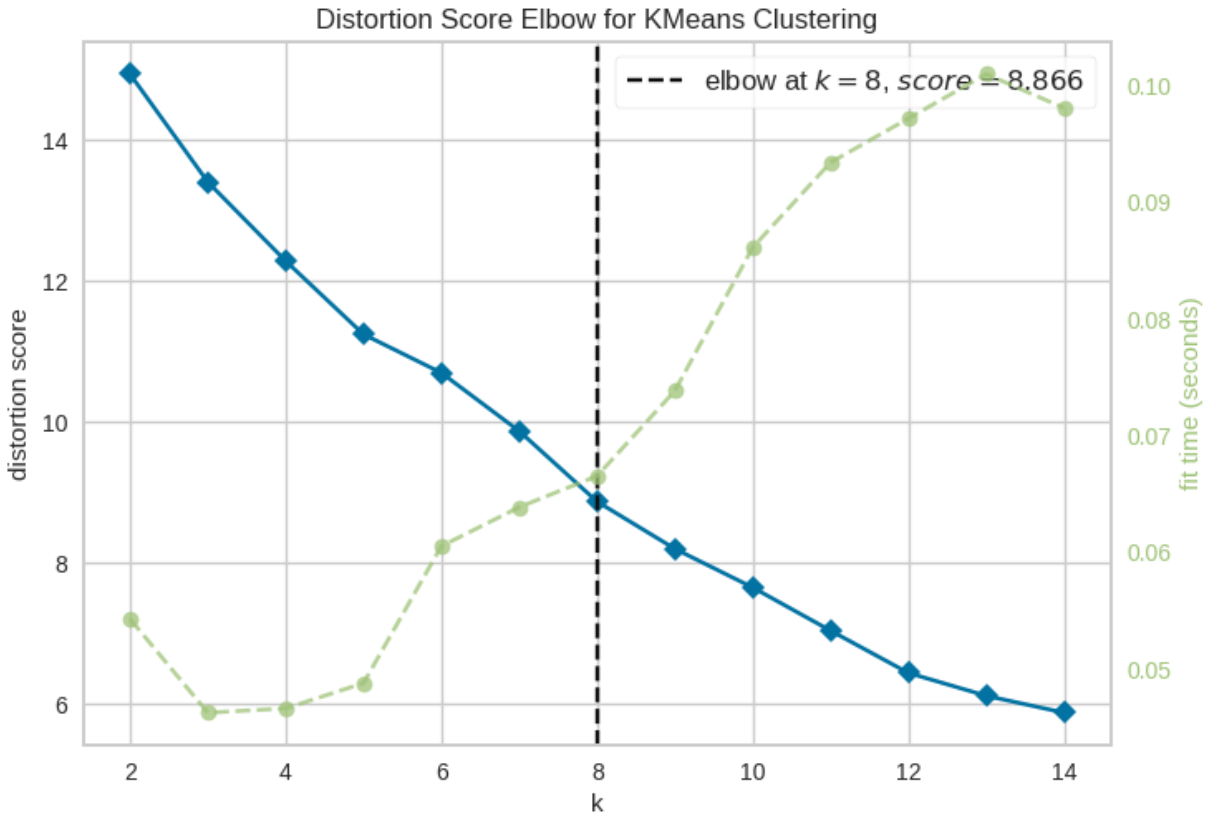
We analyze each neighborhood by taking the mean of the frequency of occurrence of each category:

	Neighborhood	Yoga Studio	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Art Gallery	Asian Restaurant	Athletics & Sports	Baby Store	Bakery	Bank
0	Bathurst Manor , Wilson Heights , Downsview No...	0.000000	0.0	0.000000	0.000000	0.000000	0.0000	0.000	0.000	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.117647
1	Bedford Park , Lawrence Manor East,Bedford Par...	0.000000	0.0	0.000000	0.000000	0.000000	0.0000	0.000	0.000	0.000	0.040000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	Birch Cliff , Cliffside West,Birch Cliff,Birch...	0.000000	0.0	0.000000	0.000000	0.000000	0.0000	0.000	0.000	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	CN Tower , King and Spadina , Railway Lands , ...	0.000000	0.0	0.000000	0.000000	0.062500	0.0625	0.125	0.125	0.125	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	Caledonia-Fairbanks,Caledonia Fairbanks,Caledo...	0.000000	0.0	0.000000	0.000000	0.000000	0.0000	0.000	0.000	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
5	Christie,Christie,Palmerston-Little Italy,Wych...	0.000000	0.0	0.000000	0.000000	0.000000	0.0000	0.000	0.000	0.000	0.000000	0.000000	0.000000	0.066667	0.066667	0.000000	0.000000
6	Church and Wellesley,Church and Wellesley,Moun...	0.014925	0.0	0.014925	0.014925	0.000000	0.0000	0.000	0.000	0.000	0.014925	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
7	Clairville , Humberwood , Woodbine Downs , Wes...	0.000000	0.0	0.000000	0.000000	0.000000	0.0000	0.000	0.000	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Now we added our three columns:

Sushi Restaurant	Tea Room	Thai Restaurant	Theater	Theme Restaurant	Toy / Game Store	Trail	Truck Stop	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant	Wine Bar	Wine Shop	Women's Store	Number of all crimes 2014 - 2020	School Score 2018-2019	Average sale price - now
0.058824	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.925550	0.897727	0.547809
0.040000	0.0	0.04	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.855103	0.704545	0.000000
0.000000	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.786791	0.318182	0.647410
0.000000	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.953169	0.397727	0.893426
0.000000	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.25	0.961041	0.011364	0.796813

Start clustering neighborhoods with k-means and k =5. Then using Elbow Method we get optimum k :



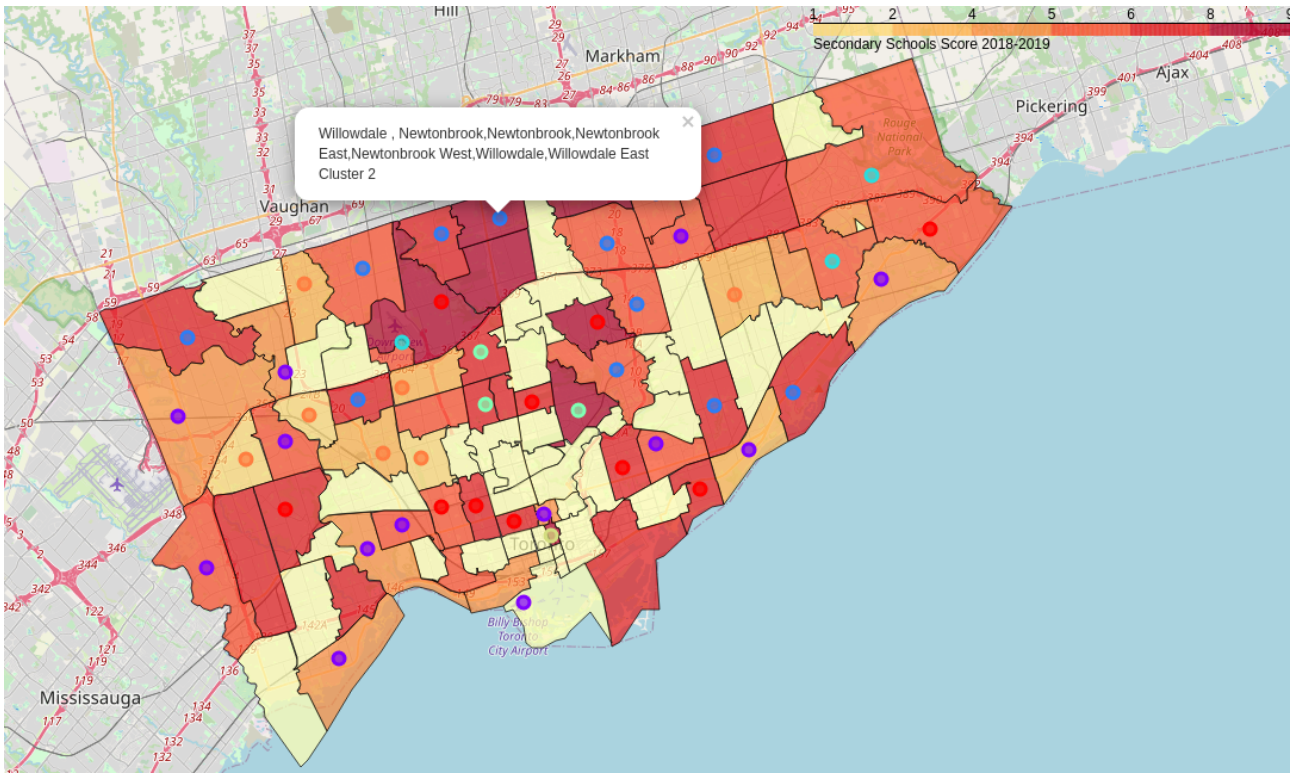
We created a data frame with all pieces of information from 'Postalcode' to Ten Most Common Venues

	Postalcode	Borough	Neighborhood	Latitude	Longitude	Number of all crimes 2014 - 2020	School Score 2018-2019	Average sale price - now	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	M3A	North York	Parkwoods,Parkwoods,Parkwoods-Donalda	43.753259	-79.329656	2014.0	6.0	1000000.0	1.0	Park	Fast Food Restaurant	Food & Drink Shop	Women's Store	Dance Studio	Discount Store	Diner	Dim Sum Restaurant	Dessert Shop	Department Store
1	M6A	North York	Lawrence Manor , Lawrence Heights,Briar Hill-B...	43.718518	-79.464763	1210.0	3.0	848000.0	7.0	Clothing Store	Furniture / Home Store	Carpet Store	Miscellaneous Shop	Boutique	Sporting Goods Shop	Coffee Shop	Accessories Store	Vietnamese Restaurant	Gastropub
2	M9A	Etobicoke	Islington Avenue,Edenbridge-Humber Valley,Humb...	43.667856	-79.532242	729.0	7.3	1800000.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	M1B	Scarborough	Malvern , Rouge,Agincourt South-Malvern West,M...	43.806686	-79.194353	3051.0	5.5	707000.0	1.0	Fast Food Restaurant	Women's Store	Donut Shop	Distribution Center	Discount Store	Diner	Dim Sum Restaurant	Dessert Shop	Department Store	Deli / Bodega
4	M3B	North York	Don MillsNorth,Banbury-Don Mills,Don Mills	43.745906	-79.352188	1402.0	8.1	1500000.0	0.0	Gym	Athletics & Sports	Baseball Field	Caribbean Restaurant	Café	Japanese Restaurant	Dog Run	Distribution Center	Discount Store	Diner
5	M5B	Downtown Toronto	Garden District,Church-Yonge Corridor,Garden D...	43.657162	-79.378937	7877.0	9.4	754000.0	2.0	Coffee Shop	Clothing Store	Sandwich Place	Café	Hotel	Bank	Japanese Restaurant	Pizza Place	Cosmetics Shop	Theater

There are two problems. The first one, Foursquare API has always updated data - some venus no longer exist. We need to check if for some neighborhoods there is no available data. If so we must drop that rows.
Second problem - a type of 'Cluster Labels' should be an integer.

	Postalcode	Borough	Neighborhood	Latitude	Longitude	Number of all crimes 2014 - 2020	School Score 2018-2019	Average sale price - now	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	M3A	North York	Parkwoods,Parkwoods,Parkwoods-Donalda	43.753259	-79.329656	2014.0	6.0	1000000.0	1	Park	Fast Food Restaurant	Food & Drink Shop	Women's Store	Dance Studio	Discount Store
1	M6A	North York	Lawrence Manor , Lawrence Heights,Briar Hill-B...	43.718518	-79.464763	1210.0	3.0	848000.0	7	Clothing Store	Furniture / Home Store	Carpet Store	Miscellaneous Shop	Boutique	Sporting Goods Shop
3	M1B	Scarborough	Malvern , Rouge,Agincourt South-Malvern West,M...	43.806686	-79.194353	3051.0	5.5	707000.0	1	Fast Food Restaurant	Women's Store	Donut Shop	Distribution Center	Discount Store	Diner
4	M3B	North York	Don MillsNorth,Banbury-Don Mills,Don Mills	43.745906	-79.352188	1402.0	8.1	1500000.0	0	Gym	Athletics & Sports	Baseball Field	Caribbean Restaurant	Café	Japanese Restaurant
5	M5B	Downtown Toronto	Garden District,Church-Yonge Corridor,Garden D...	43.657162	-79.378937	7877.0	9.4	754000.0	2	Coffee Shop	Clothing Store	Sandwich Place	Café	Hotel	Bank

Visualization of the resulting clusters on the map:



We see, according to this clusterization - the most optimum place to settle down in Toronto would be neighborhoods from cluster 5. Secondary schools with high score 7.3 to 8.1, number of crimes 739 to 2108, avg house prices $ 791000 to $ 1200000 and 1st Most Common Venue is Park for a family with children.

Cluster 5

```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 4, toronto_merged.columns[[1] +[2]+ list(range(5, toronto_merged.shape[1]))]]
```

	Borough	Neighborhood	Number of all crimes 2014 - 2020	School Score 2018-2019	Average sale price - now	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
21	East York/East Toronto	The Danforth East,Blake-Jones,Broadview North...	739.0	7.3	1300000.0	4	Park	Metro Station	Convenience Store	Women's Store	Dance Studio	Discount Store	Diner	Dim Sum Restaurant
26	North York	Willowdale , Newtonbrook,Newtonbrook,Newtonbro...	1110.0	8.1	918000.0	4	Park	Women's Store	Curling Ice	Distribution Center	Discount Store	Diner	Dim Sum Restaurant	Dessert Shop
43	Scarborough	Milliken , Agincourt North , Steeles East , L'...	2108.0	7.5	898000.0	4	Park	Playground	Intersection	Curling Ice	Discount Store	Diner	Dim Sum Restaurant	Dessert Shop

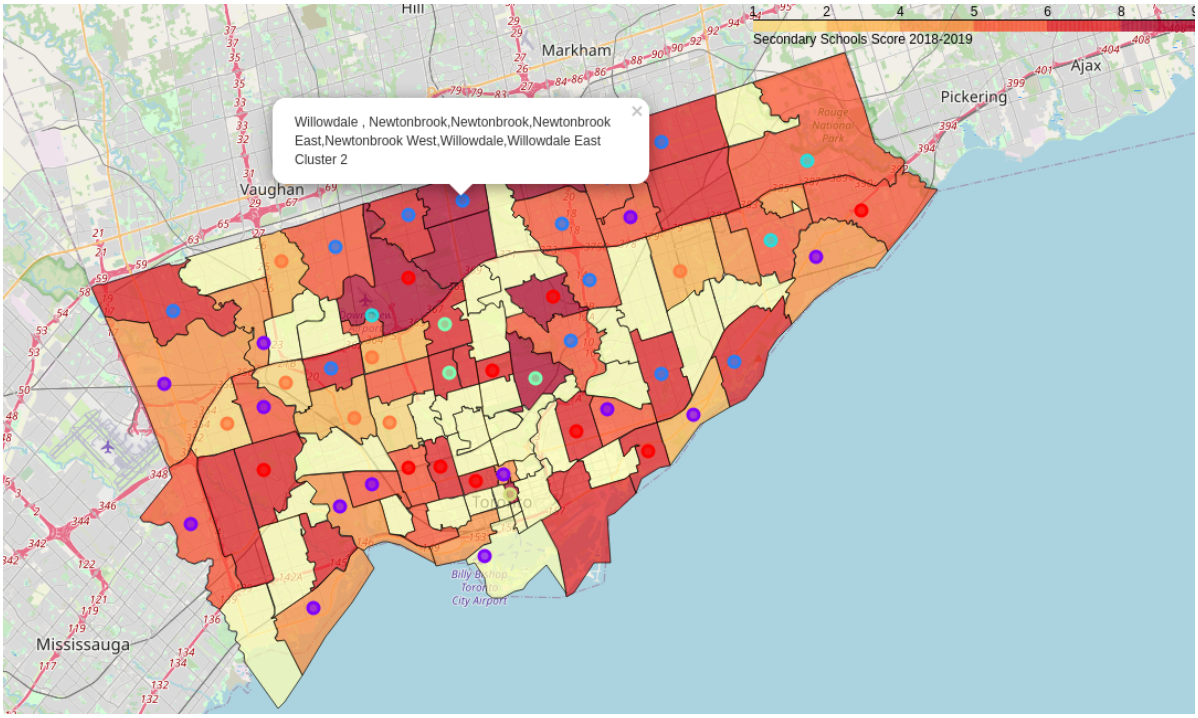
Results:

Two clusterings - first only with 'Number of all crimes 2014 - 2020','School Score 2018-2019','Average sale price - now' and second with data frame has all venus shows "similar" results.

Although the number of clusters was different - the first one has k=7 and the second k=8.

Discussion:

The most important observation was the problem with getting all the needed information - it's really hard. Sometimes there is no data or data are a few years old.



The second observation of the decision-making process on data - where to settle down on certain conditions is a tough one. We never have full information, and data are always changing. Data acquiring from yesterday, today appears not exact. So our predictions sometimes are similar to weather predictions.

For example, venues from Foursquare API, few hours - change the number of all venues in many categories and neighborhoods in Toronto. What changes our data.

The main result is in agreement with our intuition. It is impossible to choose the best optimum, but some results - from North York borough and "Willowdale, Newtonbrook, Newtonbrook, Newtonbrook East, Newtonbrook West, Willowdale, Willowdale East" neighborhoods seems perfect for a couple with a two son who wants to immigrate to Canada and settle down in Toronto.

Conclusion

Our data science tools in python such as web scraping using BeautifulSoup, folium, mathplotlib, pandas, and machine learning help people - in this case, immigrants families - to choose a better place to settle down and live with children. Nowadays these tools help other people from all over the world choose the right place if they are forced to leave their home and country. Data science simplifies looking for such things as a good school, secure neighborhood, price of a house, and place for spending outdoor time, which in older times seemed just as a "lottery".