The battle of Neighborhood

Capstone project

Chennoufi Mohamed Soheib

4/28/20

Data Science

1. Introduction/Business Problem

Discussion of the business problem and the audience who would be interested in this project.

1.1 Background

The average Canadian moves about eleven times in their lifetime. This brings us to the question: Do people move until they find a place to settle down where they truly feel happy, or do our wants and needs change over time, prompting us to eventually leave a town we once called home for a new area that will bring us satisfaction ? or, do we too often move to a new area without knowing exactly what we're getting into, forcing us to turn tail and run at the first sign of discomfort?

To minimize the chances of this happening, we should always do proper research when planning our next move in life. Consider the following factors when picking a new place to live so you do not end up wasting your valuable time and money making a move you'll end up regretting. Safety is a top concern when moving to a new area. If you don't feel safe in your own home, you're not going to be able to enjoy living there.

1.2 Problem

The crime statistics dataset of TORONTO found has crimes in each Neighborhood of Toronto from 2014 to 2019. The year 2019 being the latest, we will be considering the data of that year which is actually old information as of now. The crime rates in each borough may have changed over time.

This project aims to select the safest borough in Toronto based on the total crimes, explore the neighborhoods to find the 10 most common venues in each neighborhood and finally cluster the neighborhoods using k-mean clustering.

1.3 Interest

Expats who are considering relocating to Toronto will be interested to identify the safest Borough in Toronto, analyze the Neighborhoods there and explore the common venues around each one.

2. Data Acquisition and cleaning:

2.1 Data Acquisition

The data acquired for this project is a combination of data from three sources.

The first data source of the project uses a Toronto crime data that shows the crime per Neighborhood in Toronto this data is from the Toronto Polish Department Data http://data.torontopolice.on.ca/datasets/mci-2014-to-2019. I made changes to remove the useless data. The dataset contains the following columns:

- Neighborhood: the common name of Toronto.
- Population : the population of each neighborhood
- AVG : the average of crimes for each category (between 2014 and 2019)
- Total: the total average of crimes of all categories.

The second source of data is scraped from a Wikipedia page that contains the list of Toronto Neighborhoods. The columns are : Borough and Neighborhood.

- Borough: The names of the 10 Boroughs of Toronto.
- Neighborhood: the names of Neighborhood of each Borough.

The third source of data is the one that we used in the last assignment which contains the list of all the Neighborhoods, Boroughs and their Latitude and Longitude.

2.2 Data Cleaning

The data preparation for each of the three sources of data is done separately.

1/- From the Toronto crime data, the crimes during the years (2014-2019) are selected.

	Neighbourhood	Hood_ID	Population	Assault_AVG	AutoTheft_AVG	BreakandEnter_AVG	Homicide_AVG	Robbery_AVG	TheftOver_AVG	total
0	Waterfront Communities-The Island	77	65913	851.8	53.7	247.3	1.0	82.2	56.2	1292.2
1	Bay Street Corridor	76	25797	771.0	32.8	158.7	1.5	121.3	52.3	1137.6
2	Church-Yonge Corridor	75	31340	642.8	37.8	188.5	2.0	135.7	33.8	1040.6
3	West Humber-Clairville	1	33312	301.8	366.7	137.8	1.5	91.8	52.2	951.8
4	Moss Park	73	20506	474.7	30.2	148.5	2.5	125.5	18.8	800.2
5	York University Heights	27	27593	333.2	106.3	113.2	0.8	75.8	36.3	665.6
6	Downsview-Roding-CFB	26	35052	395.8	107.8	78.8	1.3	64.7	15.2	663.6
7	Kensington-Chinatown	78	17945	368.2	27.5	150.8	1.5	64.0	26.7	638.7
8	Woburn	137	53485	384.7	46.0	105.2	1.2	83.5	13.7	634.3
9	West Hill	136	27392	402.0	26.5	82.5	0.8	65.2	6.7	583.7

2/-The second Data contains the list of Boroughs and their Neighborhoods.
It is scraped from Wikipedia page (https://en.wikipedia.org/wiki/List_of_city-designated_neighbourhoods_in_Toronto) using the Beautiful soup library in Python. Using this library, we can extract the data in the table as shown in the webpage.

After the web scraping, string manipulation is required to get the names of boroughs in the correct form. It is important because we will merge the two datasets using the Neighborhood names.

	Borough	Neighbourhood
0	Scarborough	Agincourt
1	Scarborough	Agincourt
2	Etobicoke	Alderwood
3	Old City of Toronto	The Annex
4	North York	Don Mills
5	North York	Bathurst Manor
6	Old City of Toronto	Bay Street
7	North York	Bayview Village
8	North York	Bayview Woods
9	North York	Bedford Park

The two datasets are merged on the Neighborhood names to form a new dataset that combines the necessary information in one dataset.
The purpose is to visualize the crimes rates in each Borough with the average crime number from 2014 to 2019.

	Borough	Population	Assault_AVG	AutoTheft_AVG	BreakandEnter_AVG	Homicide_AVG	Robbery_AVG	TheftOver_AVG	total
0	Scarborough	259450	1635.9	306.2	584.5	5.6	354.4	70.3	2956.9
1	North York	162689	811.6	263.7	315.2	3.1	157.8	63.2	1614.6
2	Etobicoke	89067	550.1	112.1	193.6	2.2	71.9	31.5	961.4
3	Old City of Toronto	82751	527.2	76.5	233.8	2.1	74.5	33.2	947.3
4	York	34803	225.0	40.2	50.5	1.5	48.1	5.2	370.5
5	East York	27699	138.6	12.9	57.9	0.9	23.1	6.9	240.3

After visualizing the crime in each borough we can find the one with the lowest crime number and hence tag that borough as the safest one.

3/-The third source of data is acquired from the list of neighborhoods that we worked with in WEEK03 assignment (Wikipedia).
This dataset is created from scratch, the pandas data frame is created with the names of neighborhood and Boroughs, latitude and longitude.

	Postal Code	Borough	Neighbourhood
0	M1B	Scarborough	Malvern
1	M1B	Scarborough	Rouge
2	M1C	Scarborough	Rouge Hill
3	M1C	Scarborough	Port Union
4	M1C	Scarborough	Highland Creek
5	M1E	Scarborough	Guildwood
6	M1E	Scarborough	Morningside
7	M1E	Scarborough	West Hill
8	M1G	Scarborough	Woburn
9	M1H	Scarborough	Cedarbrae
10	M1J	Scarborough	Scarborough Village

The coordinates of the neighborhoods is to be obtained using **Google Maps API Geocoding** to get the final dataset.

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Malvern	43.806686	-79.194353
1	M1B	Scarborough	Rouge	43.806686	-79.194353
2	M1C	Scarborough	Rouge Hill	43.784535	-79.160497
3	M1C	Scarborough	Port Union	43.784535	-79.160497
4	M1C	Scarborough	Highland Creek	43.784535	-79.160497
5	M1E	Scarborough	Guildwood	43.763573	-79.188711
6	M1E	Scarborough	Morningside	43.763573	-79.188711
7	M1E	Scarborough	West Hill	43.763573	-79.188711
8	M1G	Scarborough	Woburn	43.770992	-79.216917
9	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

The new dataset is used to generate the venues for each neighborhood using the Foursquare API.

3.Methodology

3.1 Exploratory Data Analysis

3.1.1 Statistical summary of crimes

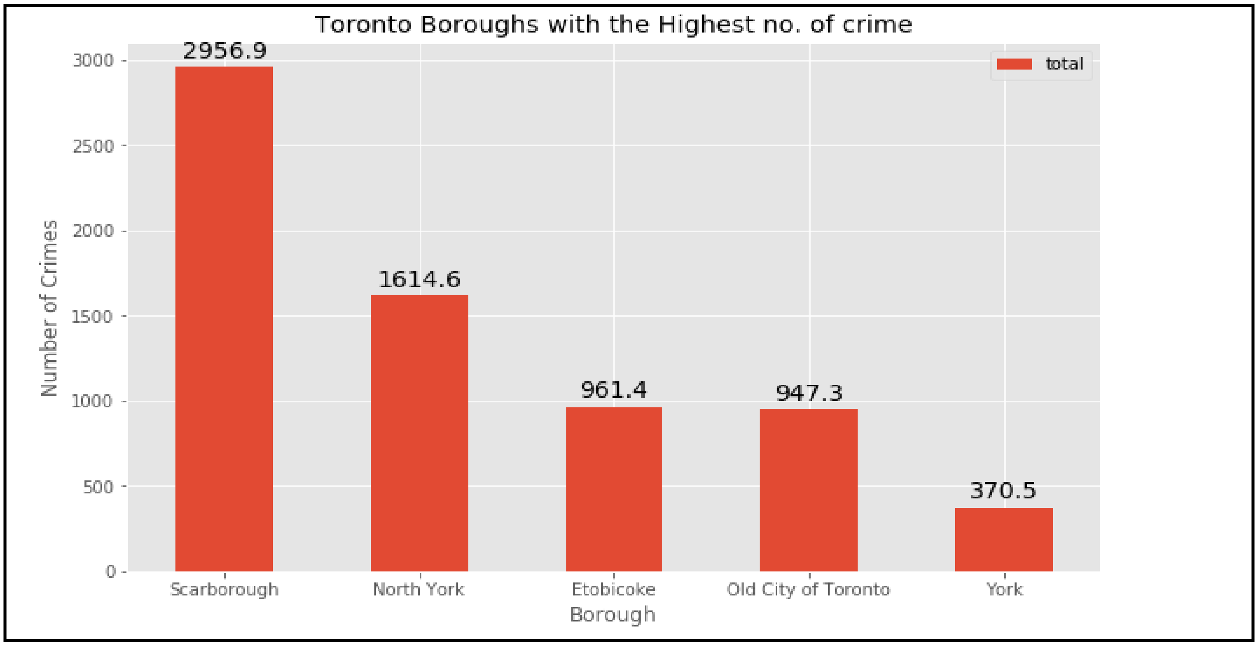
The describe function in python is used to get statistics of Toronto crimes data, this returns the mean, standard deviation, minimum, maximum, 1st quartile (25%), 2nd quartile (50%), and the 3rd quartile (75%) for each of the major categories of crime.

	Population	Assault_AVG	AutoTheft_AVG	BreakandEnter_AVG	Homicide_AVG	Robbery_AVG	TheftOver_AVG	total
count	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000
mean	109409.833333	648.066667	135.266667	239.250000	2.566667	121.633333	35.050000	1181.833333
std	87997.824224	541.351227	121.420602	197.784719	1.658513	122.718600	27.329307	998.648306
min	27699.000000	138.600000	12.900000	50.500000	0.900000	23.100000	5.200000	240.300000
25%	46790.000000	300.550000	49.275000	91.825000	1.650000	54.050000	13.050000	514.700000
50%	85909.000000	538.650000	94.300000	213.700000	2.150000	73.200000	32.350000	954.350000
75%	144283.500000	746.225000	225.800000	294.850000	2.875000	136.975000	55.700000	1451.300000
max	259450.000000	1635.900000	306.200000	584.500000	5.600000	354.400000	70.300000	2956.900000

The count for each of the major categories of crime returns the value 6, which is the number of Toronto boroughs. “Assault” is the highest reported crime during the years (from 2014 to 2019) followed by “Break and enter” and “Auto Theft”.
The lowest recorded crimes are “Homicide” and ‘Theft over’.

3.1.2 Boroughs with the highest crime rates

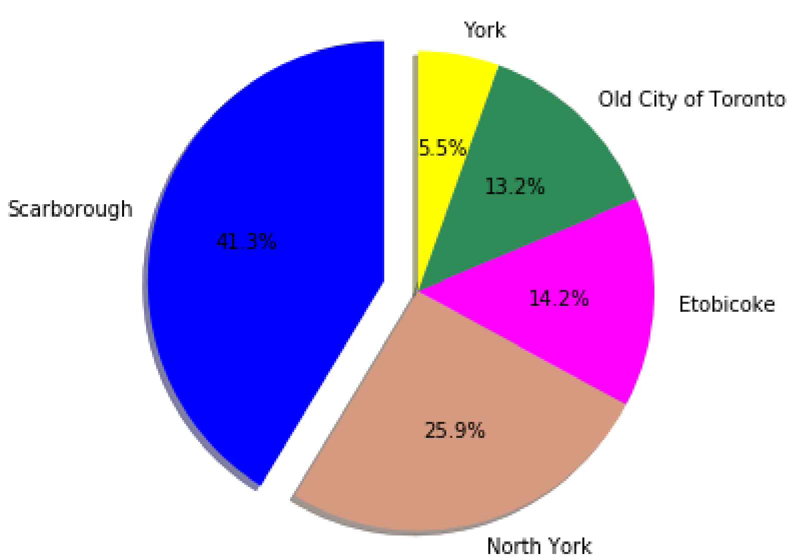
Comparing five boroughs with the highest crime rate during these years it is evident that **Scarborough** has the highest crimes recorded followed by North York, ETOBICOKE, Old city of Toronto and York.



3.1.3 Visualization of the Population

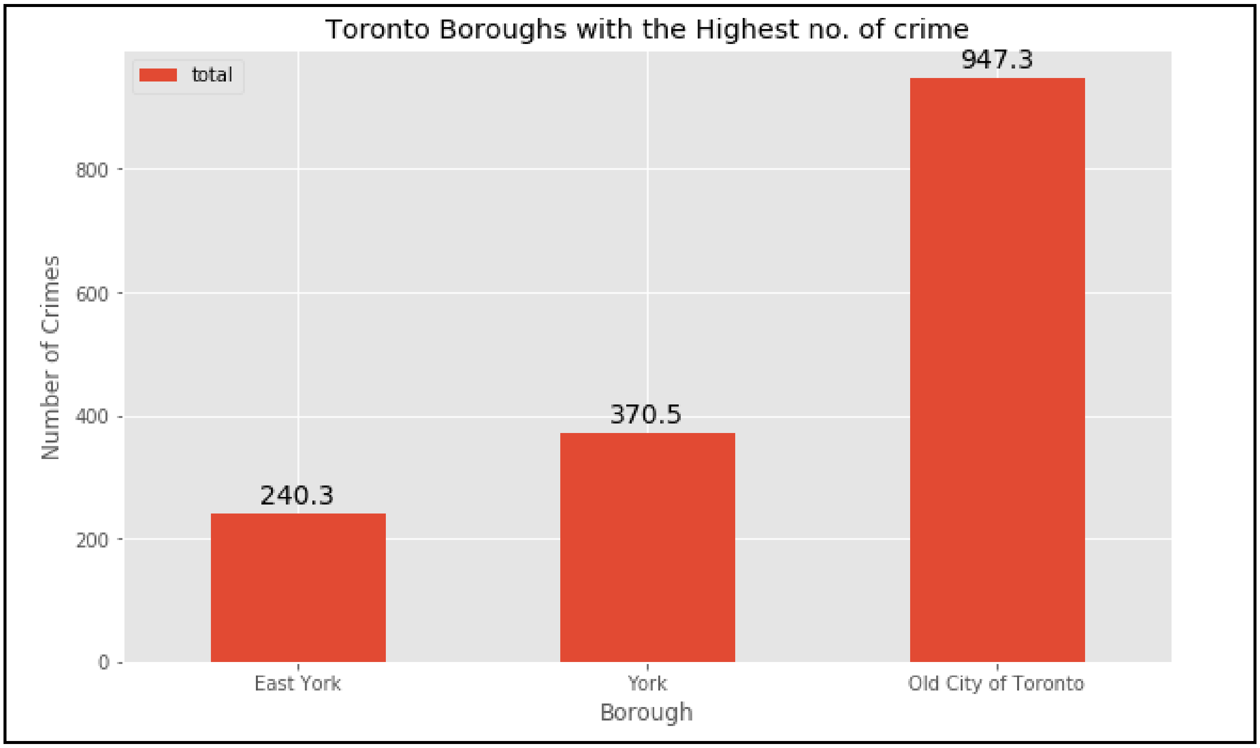
We can see that Scarborough had the highest number of population followed by North York.

We can say that there is a relation between the number of crimes and the population.

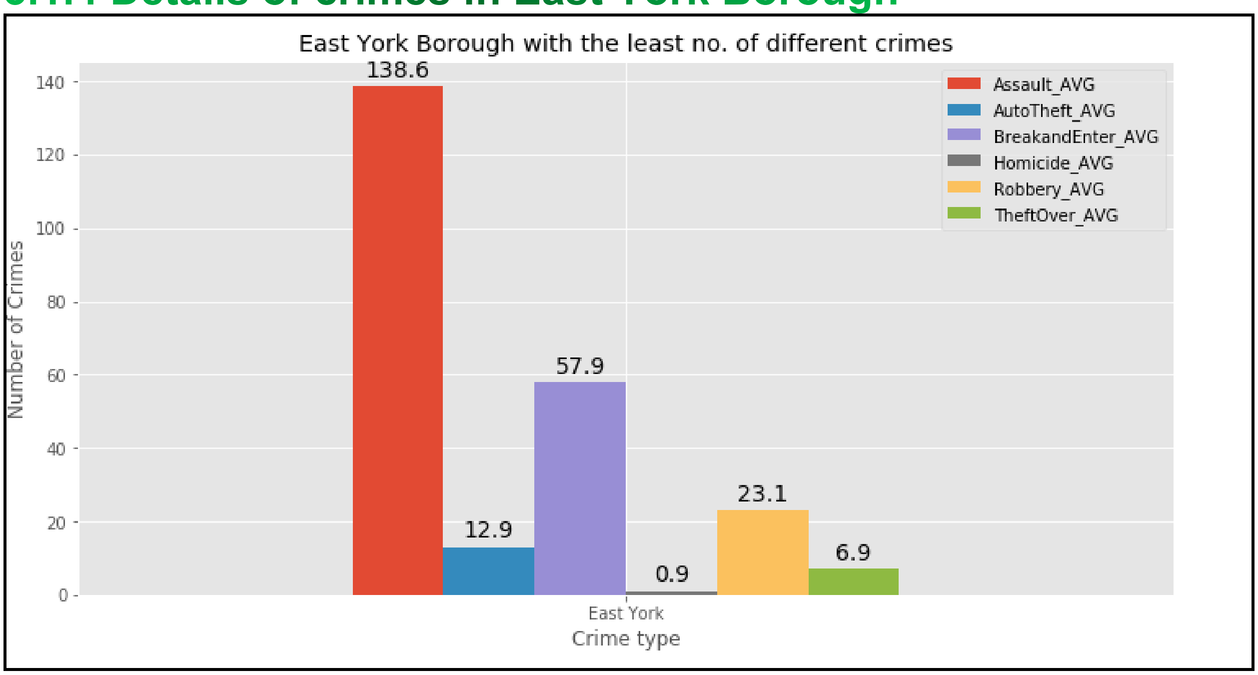


3.1.3 Boroughs with the lowest crime rates

Comparing three boroughs with the lowest crime rate during the period (2014-2019), East York has the lowest recorded crimes followed by York.



3.1.4 Details of crimes in East York Borough



Next, we will analyze data and neighborhoods in the two safest boroughs: East York and York.

3.1.4 Neighborhoods in York and East York

There are 13 neighborhoods in York and East York; they are visualized on a map using folium on python.

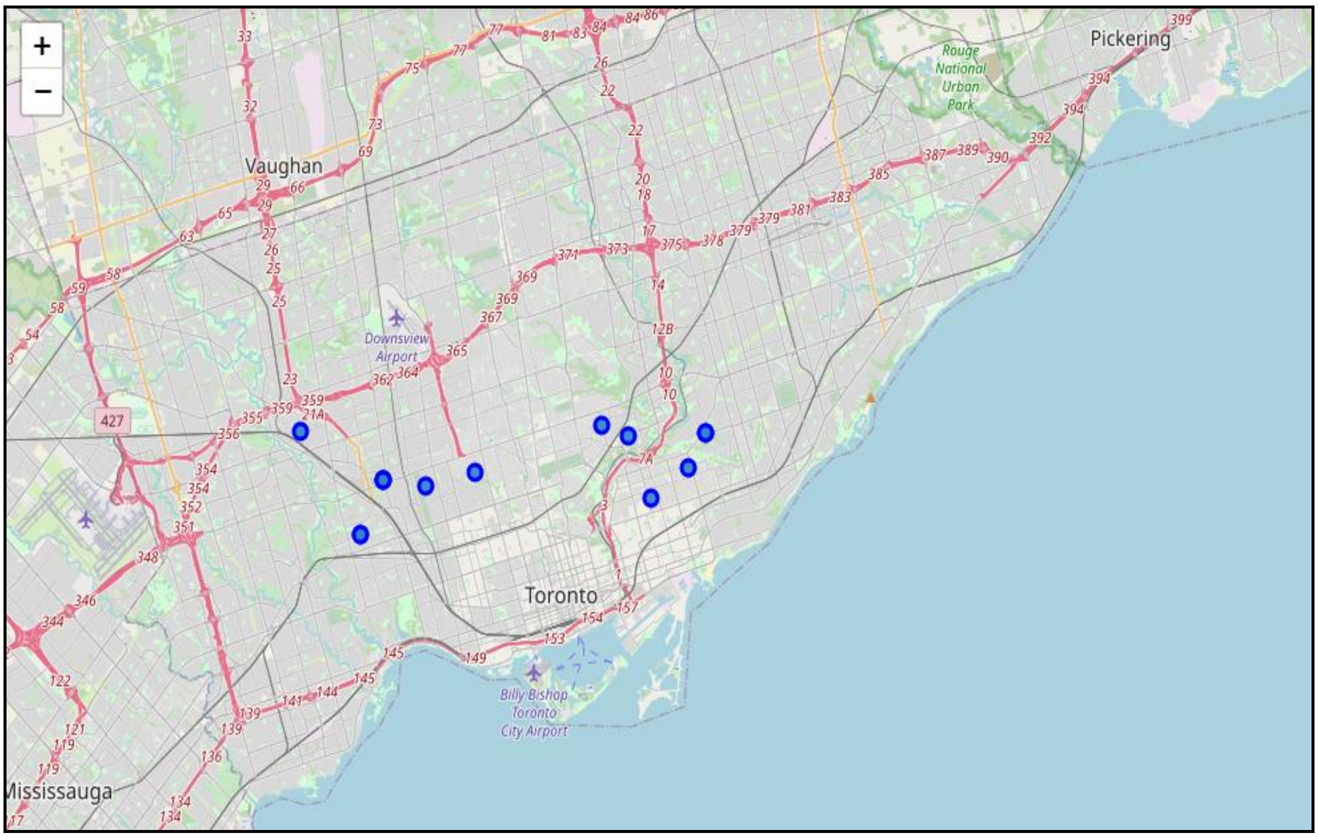


Fig 1: Neighborhoods in York and East York

3.2 Modelling

Using the final dataset containing the neighborhoods in York and East York along with the latitude and longitude, we can find all the venues within a 500 meters radius of each neighborhood by connecting to the Foursquare API.

This returns a json file containing all the venues in each neighborhood, which is converted to a pandas dataframe. This data frame contains all the venues along with their coordinates and category.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkview Hill	43.706397	-79.309937	Jawny Bakers	43.705783	-79.312913	Gastropub
1	Parkview Hill	43.706397	-79.309937	Toronto Climbing Academy	43.709362	-79.315006	Rock Climbing Spot
2	Parkview Hill	43.706397	-79.309937	Muddy York Brewing Co.	43.712362	-79.312019	Brewery
3	Parkview Hill	43.706397	-79.309937	Peek Freans Cookie Outlet	43.713260	-79.308063	Bakery
4	Parkview Hill	43.706397	-79.309937	East York Gymnastics	43.710654	-79.309279	Gym / Fitness Center
5	Parkview Hill	43.706397	-79.309937	Shoppers Drug Mart	43.705933	-79.312825	Pharmacy
6	Parkview Hill	43.706397	-79.309937	TD Canada Trust	43.705740	-79.312270	Bank
7	Parkview Hill	43.706397	-79.309937	Pizza Pizza	43.705159	-79.313130	Pizza Place
8	Parkview Hill	43.706397	-79.309937	Tim Hortons	43.714401	-79.307356	Coffee Shop
9	Parkview Hill	43.706397	-79.309937	Harvey's	43.710964	-79.309085	Fast Food Restaurant
10	Parkview Hill	43.706397	-79.309937	Nostalgia	43.706833	-79.311783	Café
11	Parkview Hill	43.706397	-79.309937	East York Animal Clinic	43.705921	-79.312196	Pet Store
12	Parkview Hill	43.706397	-79.309937	Rise & Dine Eatery	43.705769	-79.311638	Breakfast Spot
13	Parkview Hill	43.706397	-79.309937	St. Clair Ave E & O'Connor Dr	43.705233	-79.313274	Intersection
14	Parkview Hill	43.706397	-79.309937	Venice Pizza	43.705921	-79.313957	Pizza Place
15	Parkview Hill	43.706397	-79.309937	Harvey's	43.708136	-79.314105	Fast Food Restaurant

One hot encoding is done on the venues data. (One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction).

The Venues data is then grouped by the Neighborhood and the mean of the venues are calculated, finally the 10 common venues are calculated for each of the neighborhoods.

To help people find similar neighborhoods in the safest borough we will be clustering similar neighborhoods using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use a cluster size of 4 for this project that will cluster the 13 neighborhoods into 4 clusters.

The reason to conduct a K- means clustering is to cluster neighborhoods with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each neighborhood.

4. Results

After running the K-means clustering we can access each cluster created to see which neighborhoods were assigned to each of the five clusters. Looking into the neighborhoods in the first cluster .

	Borough	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
2	East York	0	Coffee Shop	Park	Pizza Place	Sandwich Place	Skating Rink	Thai Restaurant	Plaza	Diner	Pub	Bus Line
3	York	0	Pizza Place	Coffee Shop	Sushi Restaurant	Park	Bagel Shop	Bus Line	Frozen Yogurt Shop	Convenience Store	Optical Shop	Dance Studio
11	York	0	Coffee Shop	Pizza Place	Brewery	Park	Pharmacy	Gas Station	Sandwich Place	Bus Line	Beer Store	Burger Joint
12	York	0	Coffee Shop	Pizza Place	Brewery	Park	Pharmacy	Gas Station	Sandwich Place	Bus Line	Beer Store	Burger Joint

Fig 2: cluster 1

The cluster one is the biggest cluster with 4 of the 13 neighborhoods in York and East York.
Upon closely examining these neighborhoods we can see that the most common venues in these neighborhoods are Restaurants, Café. and stores.

Looking into the neighborhoods in the second, third and fourth clusters:

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

	Borough	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
3	East York	1	Sporting Goods Shop	Coffee Shop	Electronics Store	Grocery Store	Furniture / Home Store	Bank	Restaurant	Sandwich Place	Burger Joint	Sports Bar
4	East York	1	Coffee Shop	Indian Restaurant	Grocery Store	Afghan Restaurant	Brewery	Gym	Burger Joint	Shopping Mall	Supermarket	Bank

Fig 3: cluster 2

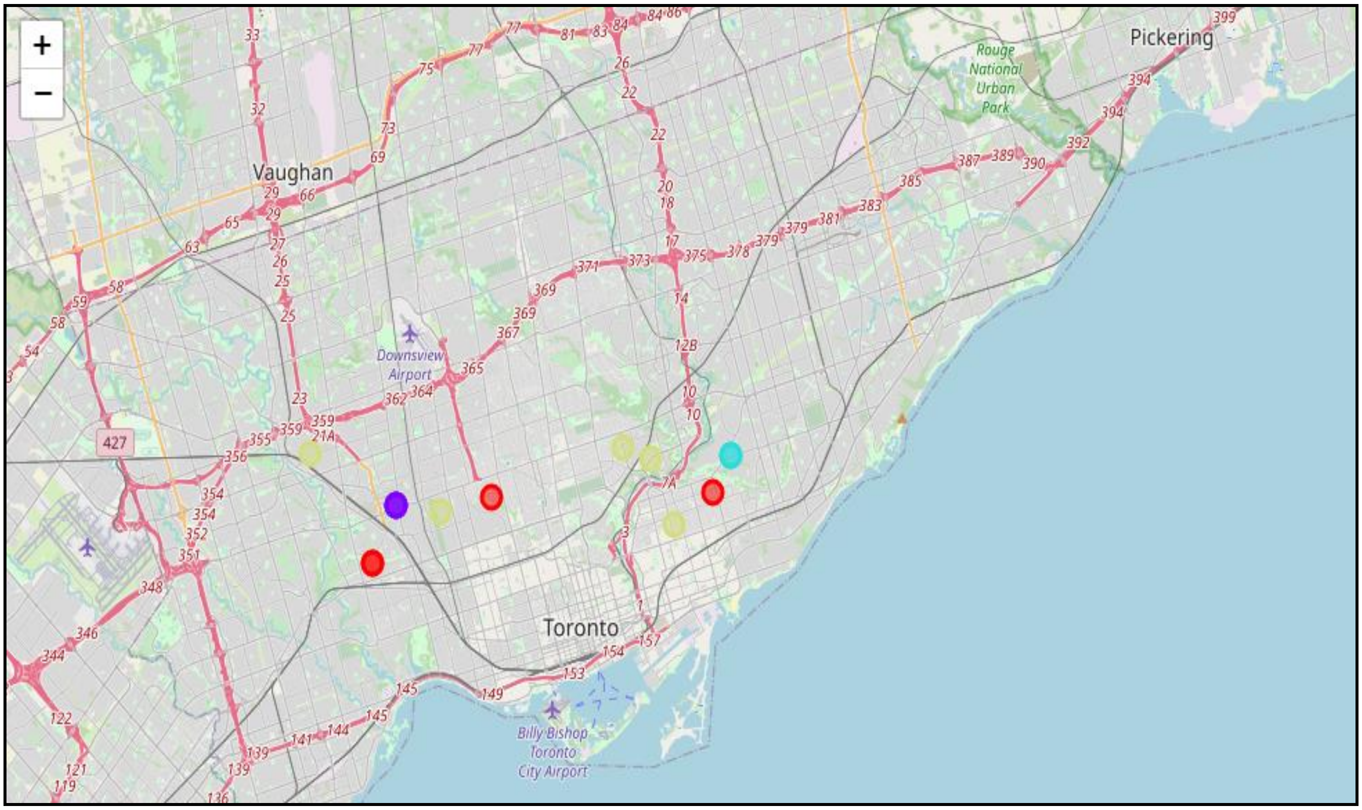
	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	East York	2	Pizza Place	Fast Food Restaurant	Coffee Shop	Brewery	Athletics & Sports	Pharmacy	Café	Rock Climbing Spot	Construction & Landscaping	Breakfast Spot
1	East York	2	Pizza Place	Fast Food Restaurant	Coffee Shop	Brewery	Athletics & Sports	Pharmacy	Café	Rock Climbing Spot	Construction & Landscaping	Breakfast Spot

Fig 4: cluster 3

	Borough	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
4	York	3	Bus Stop	Park	Pharmacy	Grocery Store	Women's Store	Hostel	Fast Food Restaurant	Japanese Restaurant	Falafel Restaurant	Mexican Restaurant
5	East York	3	Sporting Goods Shop	Coffee Shop	Electronics Store	Grocery Store	Furniture / Home Store	Department Store	Brewery	Bank	Sports Bar	Burger Joint
6	East York	3	Coffee Shop	Grocery Store	Indian Restaurant	Supermarket	Bank	Brewery	Burger Joint	Gym	Shopping Mall	Afghan Restaurant
7	East York	3	Coffee Shop	Café	Greek Restaurant	Pizza Place	Ethiopian Restaurant	Fast Food Restaurant	Beer Bar	Bar	Convenience Store	Pharmacy
13	York	3	Train Station	Coffee Shop	Pizza Place	Soccer Field	Furniture / Home Store	Fried Chicken Joint	Discount Store	Diner	Convenience Store	Pharmacy

Fig 5: Cluster 4

Visualizing the clustered neighborhoods on a map using the folium library:



5.Results and Discussion

The aim of this project is to help people who want to relocate to the safest borough in Toronto, expats can chose the neighborhoods to which they want to relocate based on the most common venues in it. For example if a person is looking for a neighborhood with good connectivity and public transportation we can see that Clusters 1 and 4 have Train stations and Bus Lines as the most common venues.

If a person is looking for a neighborhood with stores and restaurants in a close proximity then the neighborhoods in the first cluster is suitable.

For a family I feel that the neighborhoods in Cluster 2 and 3 are more suitable dues to the common venues in that cluster, these neighborhoods have common venues such as Parks, Gym centers, Bus Stops, Restaurants, Electronics Stores and Soccer fields that is ideal for a family.

6.Conclusion

This project helps a person get a better understanding of the neighborhoods with respect to the most common venues in that neighborhood.
It is always helpful to make use of technology to stay one-step ahead i.e. finding out more about places before moving into a neighborhood.
We have just taken safety as a primary concern to shortlist the borough of Toronto. The future of this project includes taking other factors such as cost of living in the areas into consideration to shortlist the borough based on safety and a predefined budget.