CAPSTONE PROJECT – THE BATTLE OF NEIGHBOURHOODS
REPORT

1.Introduction

Background:

Whenever we hear about Canada in our conversations, we tag it into two words “cool weather” and “peaceful”. Canada is obviously more than that. Is it peaceful and no crime takes place over there at all? We will study the crime data explore it and try to see pattern of crime. Covering the whole country will be too much for our scope. Instead of Canada, we will cover their most famous city, Toronto. It is the most populous city in Canada, a multicultural city, and the country’s financial and commercial centre. Its location on the north-western shore of Lake Ontario, which forms part of the border between Canada and the United States, and its access to Atlantic shipping via the St. Lawrence Seaway and to major U.S. industrial centres via the Great Lakes have enabled Toronto to become an important international trading centre. This is the reason that many immigrants choose Toronto as it gives a lot of job and business opportunities. All these factors are conducive for business to grow.

To start a business in any neighbourhood we consider factors like infrastructure of the area, neighbourhood population profile, cost of the commercial properties i.e. cost of business. Safety is never included in balance sheet but I think it can cost a business a lot if not taken into consideration. Safety should be one of the topmost priorities of business. The neighbourhood should provide a safe environment not only for business but also for customers. This is how the business can sustain in long term.

Problem:

The project aims to find the safest borough and appropriate neighbourhoods to start a business establishment like a grocery store. First it will explore and analyse the crime data for the year 2019.Second part will be to find the different types of crimes that takes place on commercial properties. Based on the crime data, explore the neighbourhood of the safest borough to find the 10 most common venues in each neighbourhood and divide them into clusters using k-mean clustering method.

Interest:

Businessman who are considering opening new establishment like grocery store in the safest borough as well as explore its neighbourhoods and common venues around each neighbourhood.

2.Data Acquisition and Cleaning

Data Acquisition:

First set of data is crime data for Toronto which is acquired from official Toronto police website. The crime data is from 2013 to 2019. Since data has many columns, mentioning only the columns which are used in the project.

Data set URL: https://data.torontopolice.on.ca/datasets/mci-2014-to-2019

The dataset has following properties which we have selected:

* **PREMISE TYPE** - Type of property where crime occurred - we will choose commercial

* **OFFENCE**- Crime type
* **YEAR** - Recorded year - 2019 year will be filtered
* **MONTH** - Recorded month
* **DAY** - Recorded day
* **HOUR** - Recorded hour
* **Hood ID** - Neighbourhood id number
* **NEIGHBOURHOOD** - Recorded neighbourhood
* Lat - Latitude of the location where crime occurred *** This used only at the later part in visualisation section
* Long - Longitude of the location where crime occurred *** This used only at the later part in visualisation section

Second Source of data was acquired using web scraping method from Wikipedia link https://en.wikipedia.org/wiki/List_of_city-designated_neighbourhoods_in_Toronto

This page has list of neighbourhoods assigned to each borough. There are around 140 neighbourhoods in Toronto. Each has been assigned to one of 6 boroughs. There is also another important column of CDN number also known as "neighbourhood id".

CDN Number: This is neighbourhood id

City designated neighbourhood: Name of the neighbourhood

Borough: Borough name to which the respective neighbourhood is assigned

Neighbourhood covered: adjoining neighbourhood area included in same borough

Map: Map of the neighbourhood

We have used only CDN number and Borough column for our project.

Third dataset is geojson file for neighbourhoods of Toronto. This was used in the visualisation section of the project. The data was acquired from website URL https://nad.carto.com/tables/neighbourhoods_toronto/public/map

Data Cleaning:

The data preparation was needed for Toronto crime dataset and neighbourhood list. For Toronto crime dataset, the crimes for the recent year i.e. 2019 year was selected. There were many columns like date of the crime occurrence, duplicate location details, longitude and latitude of the crime spot is used later for visualisation. The neighbourhood data was cleaned using split string function. The neighbourhood column content was following in the original dataset:

E.g. Malvern(132) with its hood id in the bracket. The bracket part was removed using split string and only name of the neighbourhood was kept.

(206435, 8)

	premisetype	offence	occurrencemonth	occurrenceday	occurrencedayofweek	occurrencehour	Hood_ID	Neighbourhood
0	Commercial	Assault	March	24.0	Monday	1	132	Malvern
1	Other	B&E	September	27.0	Saturday	16	76	Bay Street Corridor
2	Commercial	B&E	March	24.0	Monday	6	1	West Humber-Clairville
3	Apartment	B&E	March	24.0	Monday	15	47	Don Valley Village
4	Commercial	Robbery - Business	May	3.0	Saturday	2	90	Junction Area

Simpler names were given to the columns

```
3]: tor_crime_df.columns = ['Type', 'Offence','Month','Day','Week','Hour','HoodID','Neighbourhood']
    tor_crime_df.head()
```

3]:

	Type	Offence	Month	Day	Week	Hour	HoodID	Neighbourhood
0	Commercial	Assault	March	24.0	Monday	1	132	Malvern
1	Other	B&E	September	27.0	Saturday	16	76	Bay Street Corridor
2	Commercial	B&E	March	24.0	Monday	6	1	West Humber-Clairville
3	Apartment	B&E	March	24.0	Monday	15	47	Don Valley Village
4	Commercial	Robbery - Business	May	3.0	Saturday	2	90	Junction Area

For second data set we used beautiful soup package to do web scraping. We extracted a list of neighbourhoods and boroughs.

```
3]: url="https://en.wikipedia.org/wiki/List_of_city-designated_neighbourhoods_in_Toronto"
    html_content = requests.get(url).text
    soup = BeautifulSoup(html_content, "lxml")
    print(soup.prettify())
```

```
<!DOCTYPE html>
<html class="client-nojs" dir="ltr" lang="en">
 <head>
  <meta charset="utf-8"/>
  <title>
   List of city-designated neighbourhoods in Toronto - Wikipedia
  </title>
  <script>
   document.documentElement.className="client-js";RLCONF={"wgBreakFrames":!1,"wgSeparatorTransformTable":["",""],"
dmy","wgMonthNames":["","January","February","March","April","May","June","July","August","September","October","N
-a5bb-d2ca05a5b493","wgCSPNonce":!1,"wgCanonicalNamespace":"","wgCanonicalSpecialPageName":!1,"wgNamespaceNumber":
_in_Toronto","wgTitle":"List of city-designated neighbourhoods in Toronto","wgCurRevisionId":964711822,"wgRevision
0,"wgIsRedirect":!1,"wgAction":"view","wgUserName":null,"wgUserGroups":["*"],"wgCategories":["Articles to be merge
urhoods in Toronto"],"wgPageContentLanguage":"en","wgPageContentModel":"wikitext","wgRelevantPageName":
"List of city-designated neighbourhoods in Toronto","wgRelevantArticleId":38958245,"wgIsProbablyEditable":!0,"wgRe
```

Extracted only the table with two columns Hood ID and Borough. There are 140 neighbourhoods and 6 boroughs in Toronto.

:

	HoodID	Borough
0	129	Scarborough
1	128	Scarborough
2	20	Etobicoke
3	95	Old City of Toronto
4	42	North York
...	...	...
135	94	Old City of Toronto
136	100	Old City of Toronto
137	97	Old City of Toronto
138	27	North York
139	31	North York

140 rows × 2 columns

Hood ID was used as primary key column to merge two tables.

	HoodID	Borough	Type	Offence	Month	Day	Week	Hour	Neighbourhood
0	129	Scarborough	House	B&E	April	26.0	Saturday	16	Agincourt North
1	129	Scarborough	Outside	Robbery - Other	March	7.0	Friday	22	Agincourt North
2	129	Scarborough	Outside	Robbery With Weapon	January	24.0	Friday	20	Agincourt North
3	129	Scarborough	Other	Assault	February	26.0	Wednesday	14	Agincourt North
4	129	Scarborough	House	B&E	January	10.0	Friday	13	Agincourt North
5	129	Scarborough	House	B&E	February	7.0	Friday	15	Agincourt North
6	129	Scarborough	Commercial	B&E W'Intent	December	11.0	Thursday	0	Agincourt North
7	129	Scarborough	House	B&E	September	26.0	Friday	11	Agincourt North
8	129	Scarborough	House	Assault	September	7.0	Sunday	5	Agincourt North
9	129	Scarborough	Outside	Robbery - Swarming	August	5.0	Tuesday	19	Agincourt North

3.Methodology

Exploratory Data Analysis:

Crime pattern in Toronto:

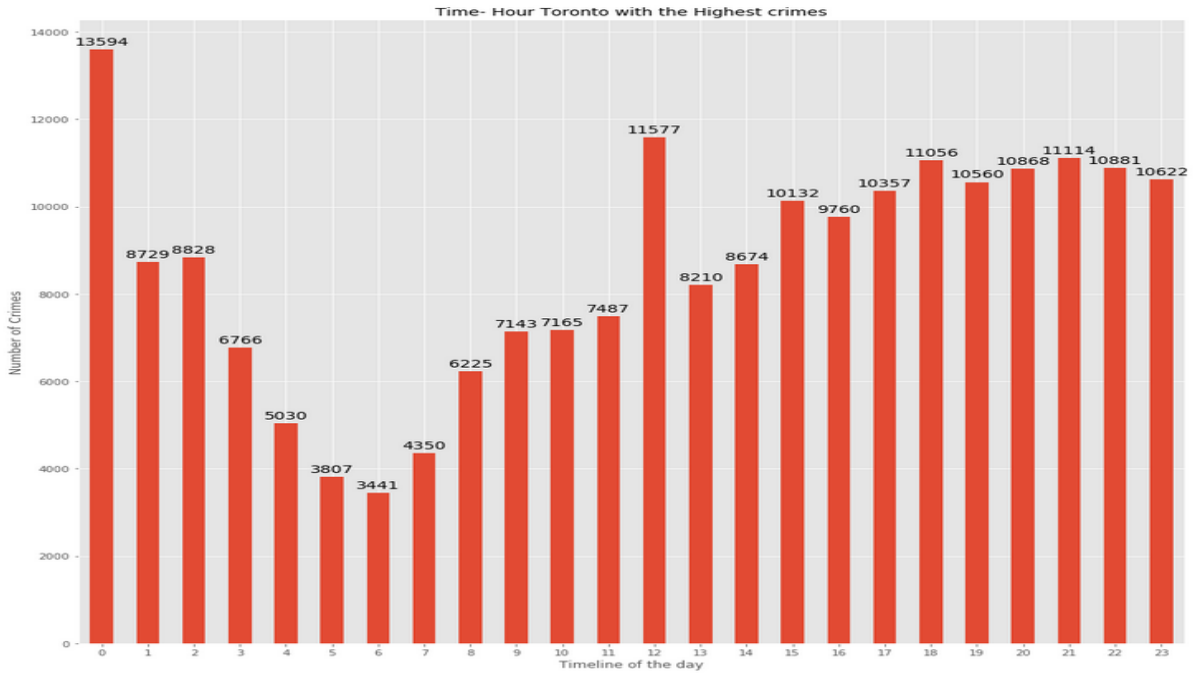
We first tried to explore and classify the count of offence by borough. Old city of Toronto has the highest number of crimes, while we can clearly see that East York and York are safest borough.

```
: Old City of Toronto    75255
  North York             45007
  Scarborough            44168
  Etobicoke              26076
  York                    9709
  East York               6161
  Name: Borough, dtype: int64
```

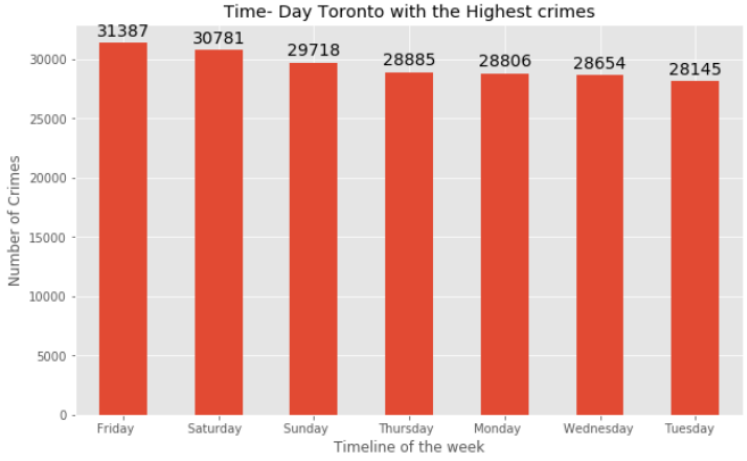
Out of these, we will see how many happen on commercial premises.

Borough	Type	
East York	Apartment	2211
	Commercial	904
	House	1185
	Other	557
	Outside	1304
Etobicoke	Apartment	5477
	Commercial	4873
	House	5603
	Other	2912
	Outside	7211
North York	Apartment	11493
	Commercial	7180
	House	10430
	Other	4612
	Outside	11292
Old City of Toronto	Apartment	17912
	Commercial	18422
	House	7565
	Other	9332
	Outside	22024
Scarborough	Apartment	9999
	Commercial	8244
	House	11301
	Other	5002
	Outside	9622
York	Apartment	2881
	Commercial	1455
	House	1816
	Other	759
	Outside	2798

Now we will see how and when these crimes generally occur so that this can help police as well as alert citizens.



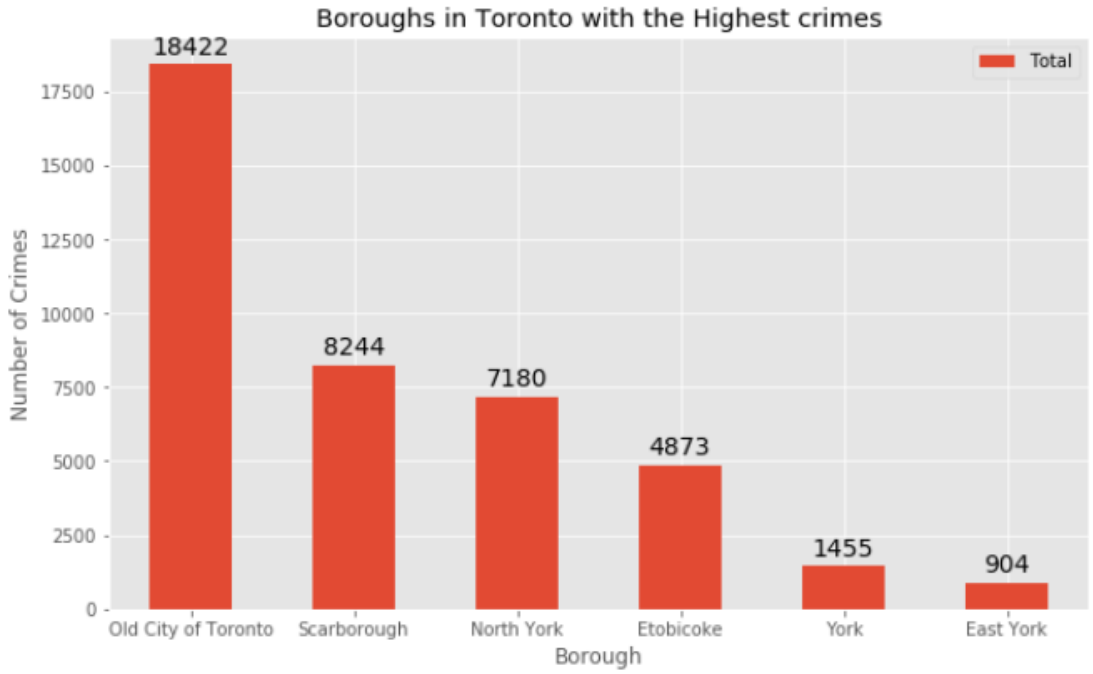
The crime is high at around evening and peaks at midnight.



The count of offence is almost same during weekdays but shoots up at weekends. People are generally out on weekends and stay usually late at night.

Using pivot, we have classified boroughs on the basis of crime and used it to plot graph

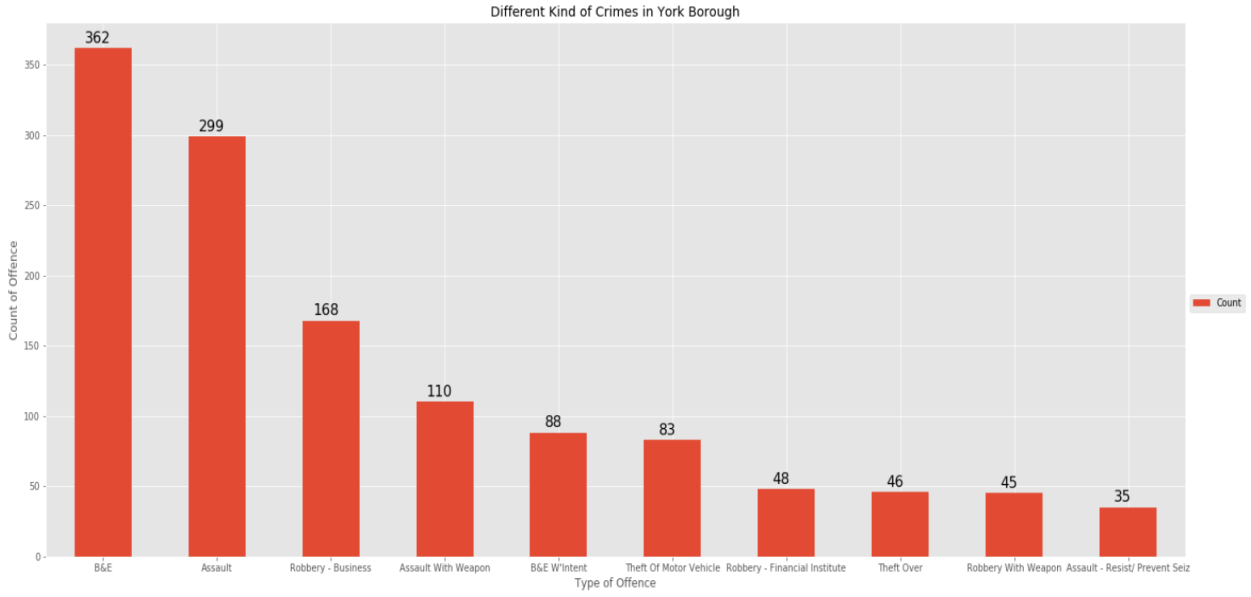
	Borough	Administering Noxious Thing	Aggravated Assault	Aggravated Assault Avails Pros	Assault	Assault - Force/Thrt /Impede	Assault - Resist/ Prevent Seiz	Assault Bodily Harm	Assault Peace Officer	Assault Peace Officer Wpn/Cbh	...	From Motor Vehicle Over	Theft Of Motor Vehicle	Theft Over	Theft Over - Bicycle	Theft Over - Distraction	Theft Over - Shoplifting	Unlawfully Causing Bodily Harm	Unlawfully In Dwelling-House	Firearm / Immit Commit Off	Total
3	Old City of Toronto	68	113	3	5642	14	765	532	312	16	...	21	382	646	8	6	173	1	8	11	18422
4	Scarborough	14	72	0	2055	5	267	149	91	5	...	57	663	219	1	2	45	0	3	5	8244
2	North York	11	65	2	1927	7	195	153	57	2	...	37	590	428	1	12	75	0	0	5	7180
1	Etobicoke	5	23	0	944	2	125	66	25	0	...	48	631	299	1	3	57	0	3	5	4873
5	York	1	13	0	299	4	35	25	9	2	...	2	83	46	0	0	5	0	0	0	1455
0	East York	0	5	0	210	2	35	15	8	0	...	2	25	38	0	0	8	0	3	0	904

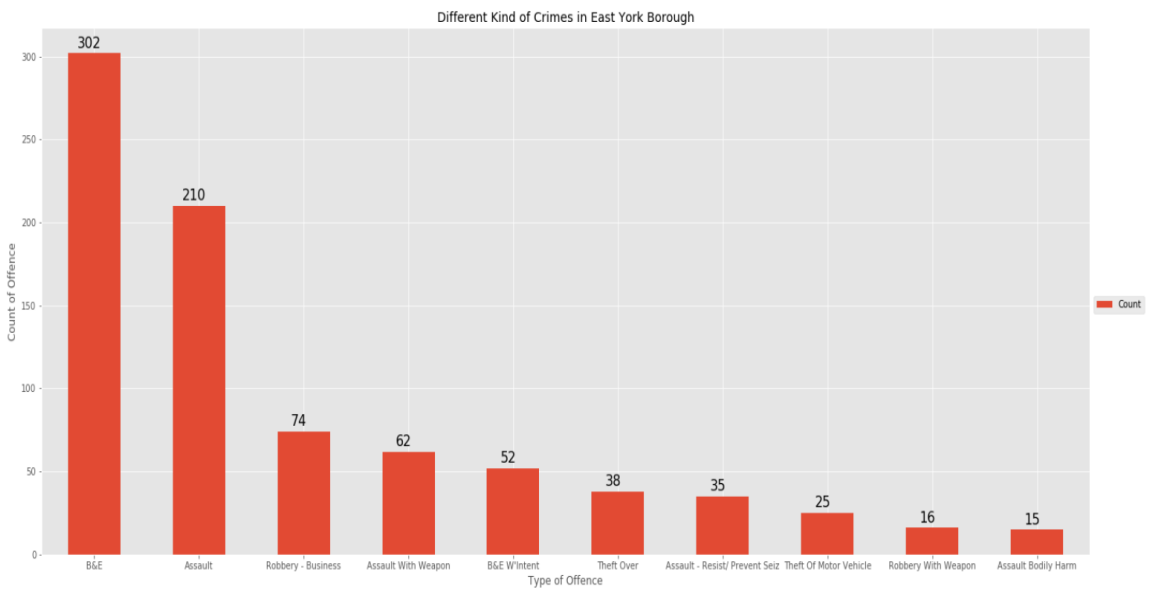


Old city of Toronto has around 88 neighbourhoods and hence the count is high, while east York and York has 8 and 10 neighbourhoods within them.

But still we would like to locate to safe borough for our commercial activity. These two areas are mostly residential areas with not very dense commercial establishments. We will consider York because it has 10 neighbourhoods, so we have more options for our locations. The crime is still on the lower side for York borough.

Comparison of crime between York vs East York

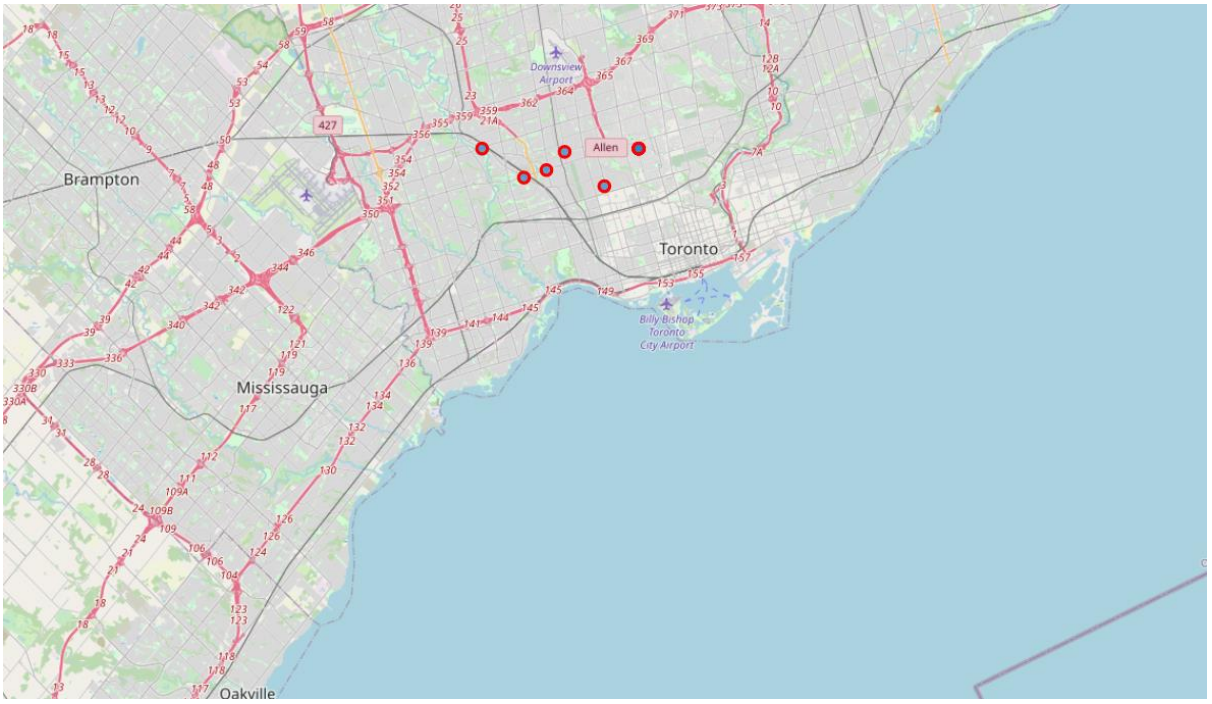




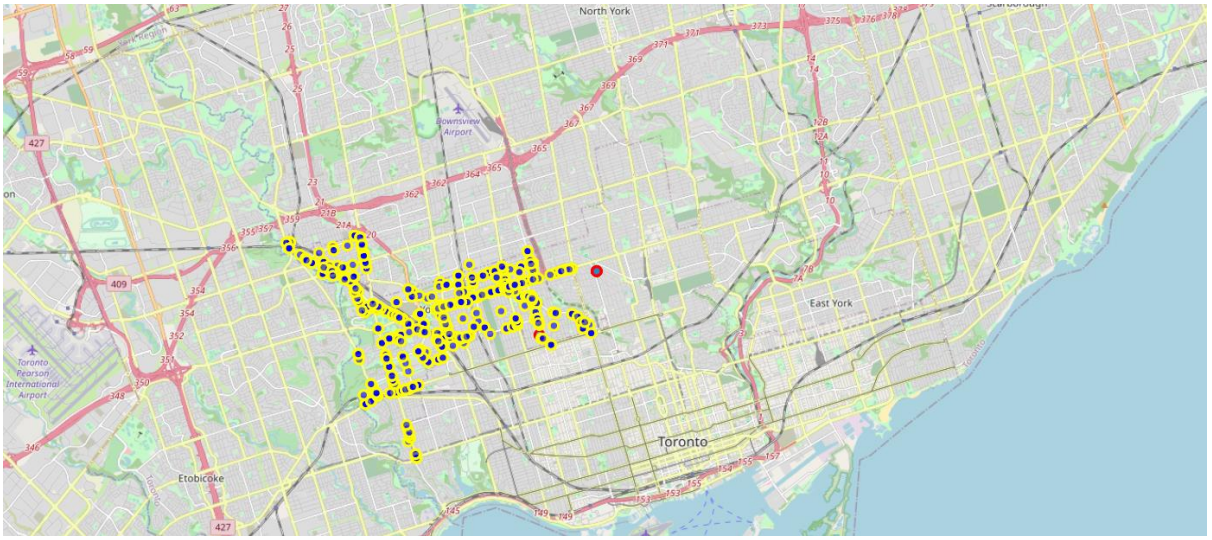
We used opencage geocoder to fetch co ordinates of neighbourhoods in East York.

	Neighbourhood	Borough	Latitude	Longitude
0	Beechborough-Greenbrook	York	43.700110	-79.416300
1	Briar Hill-Belgravia	York	43.700110	-79.416300
2	Caledonia-Fairbank	York	43.698416	-79.463480
3	Humewood-Cedarvale	York	43.700110	-79.416300
4	Keelesdale-Eglinton West	York	43.690158	-79.474998
5	Lambton Baby Point	York	43.700110	-79.416300
6	Mount Dennis	York	43.686960	-79.489551
7	Oakwood Village	York	43.682725	-79.438055
8	Rockcliffe-Smythe	York	43.700110	-79.416300
9	Weston	York	43.700161	-79.516247

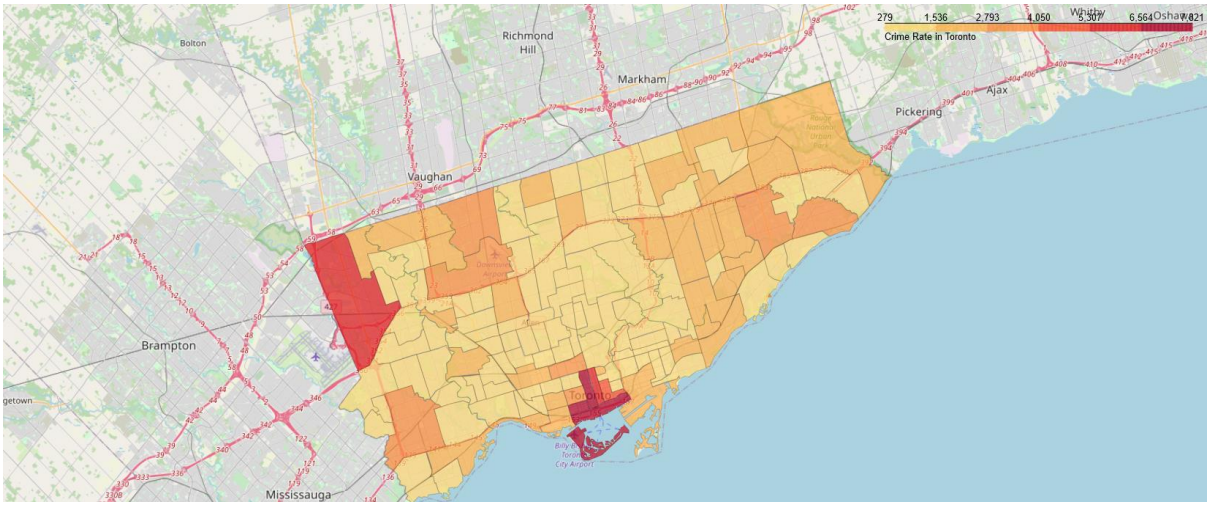
Using Folium we will plot the neighbourhoods on the map



We will use crime data to plot this on the existing map; We see that only two neighbourhoods can be seen. But let's understand what is the density of the crime in the neighbourhood



We will use choropleth map to get a full overview of the Toronto neighbourhood crime rate. Here for plotting map we used the third data set of geojson file of neighbourhoods of Toronto.



3.Modelling:

Using the final dataset along with co -ordinates details of neighbourhoods we can find all the venues within 500m radius of Toronto as center. This can be achieved by using foursquare API. It returns a geojson file which is converted into pandas dataframe.

	Neighbourhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Category
0	Beechborough-Greenbrook	43.70011	-79.4163	Hotel Gelato	Café
1	Beechborough-Greenbrook	43.70011	-79.4163	The Abbot	Gastropub
2	Beechborough-Greenbrook	43.70011	-79.4163	The Mad Bean Coffee House	Coffee Shop
3	Beechborough-Greenbrook	43.70011	-79.4163	7 Numbers	Italian Restaurant
4	Beechborough-Greenbrook	43.70011	-79.4163	The Eglinton Way	Garden

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 neighbourhoods.

	Neighbourhood	American Restaurant	Antique Shop	Asian Restaurant	BBQ Joint	Bank	Breakfast Spot	Burger Joint	Bus Line	Café	...	Sandwich Place	Skating Rink	Soccer Field	Sushi Restaurant	Taco Place	Te C
0	Beechborough-Greenbrook	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	
1	Beechborough-Greenbrook	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
2	Beechborough-Greenbrook	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
3	Beechborough-Greenbrook	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
4	Beechborough-Greenbrook	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	

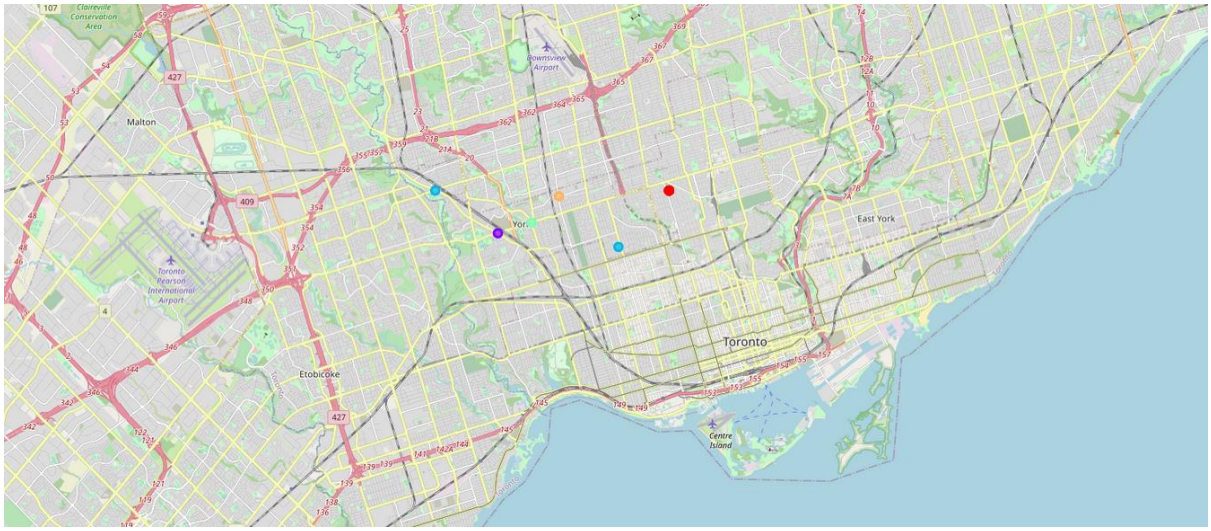
To help people find similar neighborhoods in the safest borough we will be clustering similar neighborhoods using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use a cluster size of 5 for this

project that will cluster the 10 neighborhoods into five 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 neighbourhood.

4.Results

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



Each cluster is color coded for the ease of presentation; we can see that each dot represents one cluster except cluster2.

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.

Cluster 1

```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 0, toronto_merged.columns[[1] + list(range(5, toronto_merged.shape[1]))]]
```

	Borough	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	York	Italian Restaurant	Café	Gastropub	Japanese Restaurant	Garden	Fruit & Vegetable Store	Park	Coffee Shop	Pharmacy	Sushi Restaurant
1	York	Italian Restaurant	Café	Gastropub	Japanese Restaurant	Garden	Fruit & Vegetable Store	Park	Coffee Shop	Pharmacy	Sushi Restaurant
3	York	Italian Restaurant	Café	Gastropub	Japanese Restaurant	Garden	Fruit & Vegetable Store	Park	Coffee Shop	Pharmacy	Sushi Restaurant
5	York	Italian Restaurant	Café	Gastropub	Japanese Restaurant	Garden	Fruit & Vegetable Store	Park	Coffee Shop	Pharmacy	Sushi Restaurant
8	York	Italian Restaurant	Café	Gastropub	Japanese Restaurant	Garden	Fruit & Vegetable Store	Park	Coffee Shop	Pharmacy	Sushi Restaurant

Cluster 1 has 5 neighbourhoods this is the cluster buzzing with commercial activities. The most frequent visited venue is Italian restaurants. These neighbourhood is full of restaurants, café, pubs.

This cluster looks to be the most favoured location for doing business. There is fruit and vegetable store in each neighbourhood. So we have to see that there should not be too many of them if we want to open a grocery store.

Cluster 2

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

	Borough	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
6	York	Coffee Shop	Furniture / Home Store	Pizza Place	Bus Line	Grocery Store	Tennis Court	BBQ Joint	Diner	Gift Shop	Gastropub

Cluster 3

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

	Borough	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
7	York	Pizza Place	American Restaurant	Mexican Restaurant	BBQ Joint	Bank	Breakfast Spot	Coffee Shop	Convenience Store	Dance Studio	Grocery Store
9	York	Coffee Shop	Train Station	Park	Laundromat	Middle Eastern Restaurant	Diner	Pharmacy	Pizza Place	Discount Store	Sandwich Place

Cluster 2 has grocery store, but it is 5th most common venue. It has one neighbourhood as it has unique venues that cannot be clustered with other neighbourhoods. Cluster 3 also has Restaurants but of different food cuisines. It also has convenience store while grocery store is listed at 10th place.

Cluster 4

```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 3, toronto_merged.columns[[1] + list(range(5, toronto_merged.shape[1]))]]
```

	Borough	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	Coffee Shop	Discount Store	Sandwich Place	Restaurant	Convenience Store	Gift Shop	Gastropub	Garden	Furniture / Home Store	Fruit & Vegetable Store

Cluster 5

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

	Borough	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	York	Furniture / Home Store	Coffee Shop	Burger Joint	Hardware Store	Sandwich Place	Italian Restaurant	Antique Shop	Breakfast Spot	Discount Store	Gift Shop

Similarly, we have only one neighbourhood for each cluster 4 and cluster 5. Cluster 4 has shops like discount store, gift shop, coffee shops, etc. While cluster 5 does have unique venue to its list like hardware store, burger joint, antique shop.

5.Discussion

The objective of the business problem was to help stakeholders identify one of the safest boroughs in Toronto, and an appropriate neighbourhood within the borough to set up a commercial establishment especially like a Grocery store. This has been achieved by first making use of Toronto crime data to identify a safe borough with considerable number of neighbourhood for any business to be viable. After selecting the borough it was imperative to choose the right neighbourhood where grocery shops were not among venues in a close proximity to each other. We achieved this by grouping the neighbourhoods into clusters to assist the stakeholders by providing them with relevant data about venues and safety of a given neighbourhood.

6.Conclusion

We have explored the crime data to understand different types of crimes in all neighbourhoods of Toronto and later categorized them into different boroughs, this helped us group the neighbourhoods into boroughs and choose the safest borough first. Once we confirmed the borough the number of neighbourhoods for consideration also comes down, we further shortlist the neighbourhoods based on the common venues, to choose a neighbourhood which best suits the business problem. Future scope of the project could be to take into consideration the neighbourhood profile, economic factors.