IBM Capstone Project Final Report

IBM Data Science Certification

Kevin Indra Mahendra

Report Content

1. Introduction Section
2. Data Section
3. Methodology Section
4. Results Section
5. Discussion Section
6. Conclusion Section

Introduction Section

Project Introduction

The purpose of this Project is to analyze crime incident around their venue. It will help people understanding which venue that have a high rate of crime in Toronto, Ontario, Canada.

This Project aim to create an analysis of features for a person to be carefully at venues as a comparative analysis between crimes incident place. The features include crime place, crime rates, date of incident. It will help people to anticipate possibility of crime near venues.

Problem To Be Resolved

1. Which is the closest place to a crime?
2. What categories of crime are contain in each venues?
3. What is the most common crimes on each venues?

Data Section

Data Requirement

- Toronto Crime Data from [Kaggle](). Dataset consisting of latitude and longitude, years of incident, crime category.

- Foursquare API Data about venues based on Toronto crime data.

X	Y	Index_	event_uni	occurrenc	reportedd	premisety	ucr_code	ucr_ext	offence	reportedy	reportedn	reportedd
-79.4052	43.65698	7801	GO-20152	2015-12-1	2015-12-1	Commerci	1430	100	Assault	2015	December	18
-79.3079	43.77873	7802	GO-20151	2015-08-1	2015-08-1	Commerci	1430	100	Assault	2015	August	17
-79.225	43.76594	7803	GO-20151	2015-08-1	2015-08-1	Apartmen	2120	200	B&E	2015	August	18
-79.1408	43.77865	7804	GO-20152	2015-11-2	2015-12-1	Other	2120	200	B&E	2015	December	18
-79.2884	43.69123	7805	GO-20152	2015-12-1	2015-12-1	Commerci	1430	100	Assault	2015	December	18
-79.3533	43.68855	7806	GO-20152	2015-04-1	2015-12-1	Apartmen	2120	200	B&E	2015	December	11
-79.3808	43.64989	7807	GO-20152	2015-12-1	2015-12-1	Outside	1430	100	Assault	2015	December	16
-79.3808	43.64989	7808	GO-20152	2015-12-1	2015-12-1	Outside	1430	100	Assault	2015	December	16
-79.3808	43.64989	7809	GO-20152	2015-12-1	2015-12-1	Outside	1430	100	Assault	2015	December	16
-79.3868	43.66336	7810	GO-20151	2015-11-1	2015-11-1	Outside	1610	160	Robbery -	2015	Novembe	17
-79.6113	43.71069	7811	GO-20151	2015-11-1	2015-11-1	Commerci	1480	100	Assault - F	2015	Novembe	18
-79.436	43.69877	7812	GO-20152	2015-12-0	2015-12-0	Other	1430	100	Assault	2015	December	1
-79.5086	43.72092	7813	GO-20152	2015-12-0	2015-12-0	Commerci	1430	100	Assault	2015	December	1
-79.4164	43.73579	7814	GO-20152	2015-12-0	2015-12-0	House	2120	200	B&E	2015	December	8
-79.2098	43.81421	7815	GO-20152	2015-12-0	2015-12-0	Other	1430	100	Assault	2015	December	8
-79.3598	43.65076	7816	GO-20152	2015-12-1	2015-12-1	Outside	1430	100	Assault	2015	December	11
-79.2373	43.73668	7817	GO-20152	2015-12-1	2015-12-1	Outside	1430	100	Assault	2015	December	16
-79.5653	43.73724	7818	GO-20152	2015-12-1	2015-12-1	Outside	1430	100	Assault	2015	December	16
-79.241	43.76694	7819	GO-20152	2015-12-1	2015-12-1	House	2120	200	B&E	2015	December	16
-79.3954	43.6719	7820	GO-20152	2015-12-1	2015-12-1	Commerci	2120	200	B&E	2015	December	17
-79.2616	43.81112	7821	GO-20152	2015-12-1	2015-12-1	House	2120	200	B&E	2015	December	17
-79.2168	43.76022	7822	GO-20152	2015-12-1	2015-12-1	Apartmen	1430	100	Assault	2015	December	17

Data Processing & Tools Used

- Toronto Neighborhood Crime Score data and map is to be created with use of Nominatim, Open Cage Geocode, Foursquare and Folium mapping

- Foursquare JSON Data is to be created with use of JSON Request, JSON Normalize

	Neighbourhood	Crime Score	lat	lon
0	Agincourt North	164	43.808038	-79.266439
1	Agincourt South Malvern West	219	43.795223	-79.260241
2	Alderwood	79	43.601717	-79.545232
3	Annex	371	43.670338	-79.407117
4	Banbury Don Mills	178	43.734804	-79.357243
5	Bathurst Manor	95	43.763893	-79.456367
6	Bay Street Corridor	536	43.665272	-79.387531
7	Bayview Village	125	43.769197	-79.376662
8	Bayview Woods Steeles	72	43.798127	-79.382973
9	Bedford Park Nortown	207	43.700110	-79.416300

Methodology Section

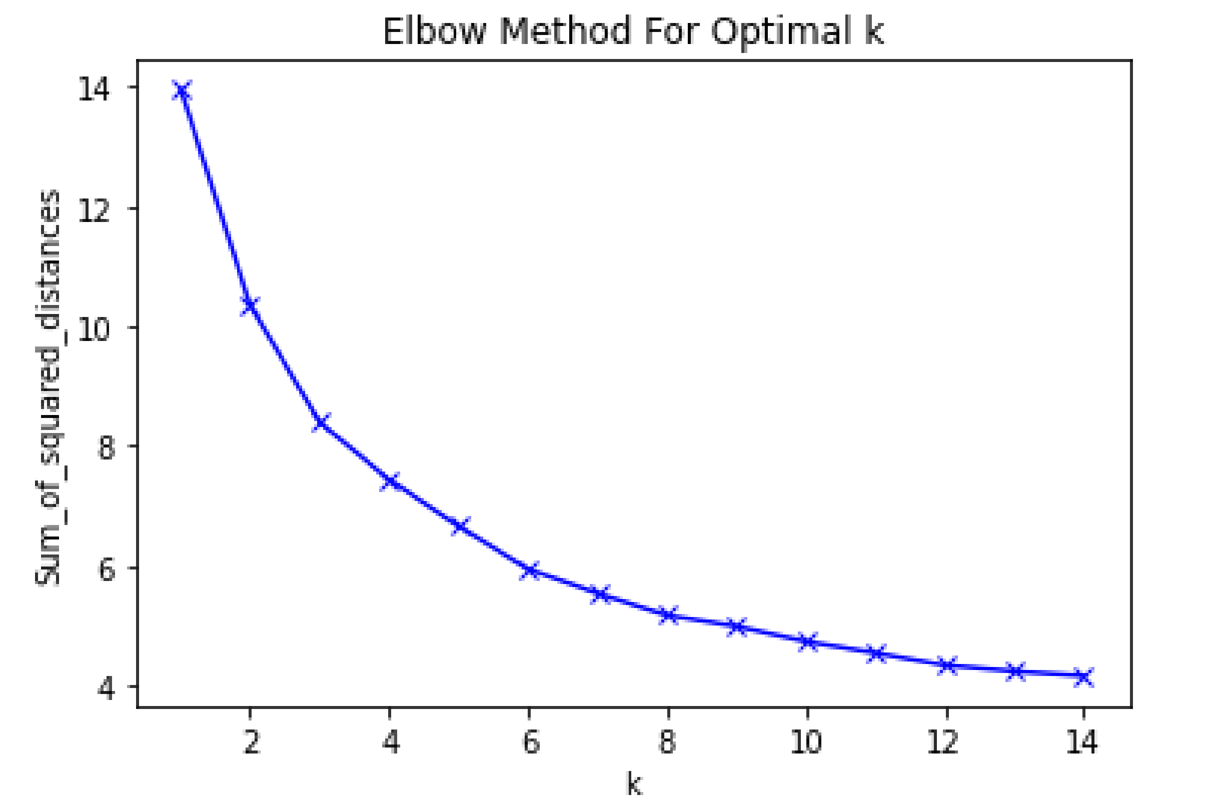
Clustering Approach

To compare the similarities of two regions, we decided to explore neighborhoods, segment them, and group them into clusters to find similar neighborhoods in a big city like New York and San Francisco. To be able to do that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm

Find Optimal K Value

For each k value, we will initialise k-means and use the inertia attribute to identify the sum of squared distances of samples to the nearest cluster centre.

As k increases, the sum of squared distance tends to zero. Imagine we set k to its maximum value n (where n is number of samples) each sample will form its own cluster meaning sum of squared distances equals zero.



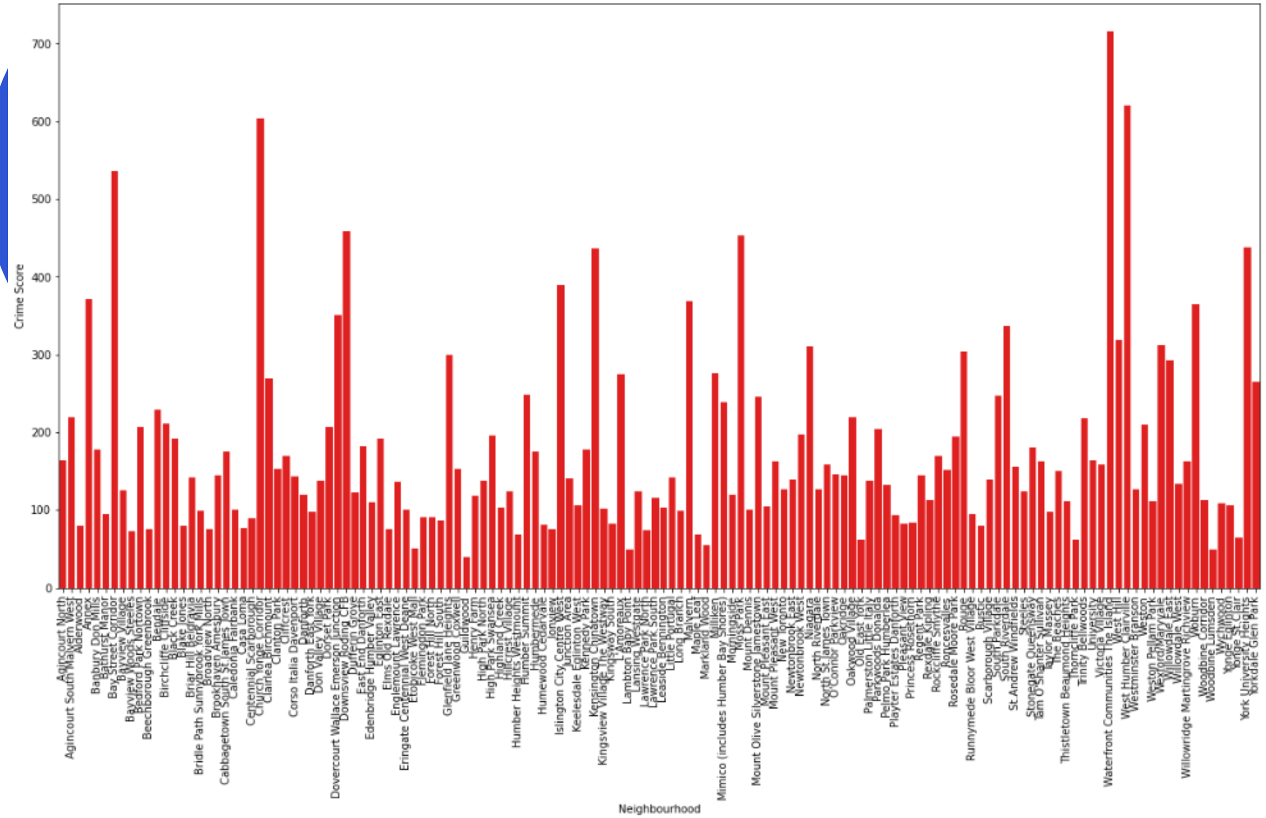
Work Flow

Using credentials of Foursquare API features of near-by places of the neighborhoods would be mined. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 1000 and the radius parameter would be set to 100.

Result & Discussion Section

Neighborhood Crime Score

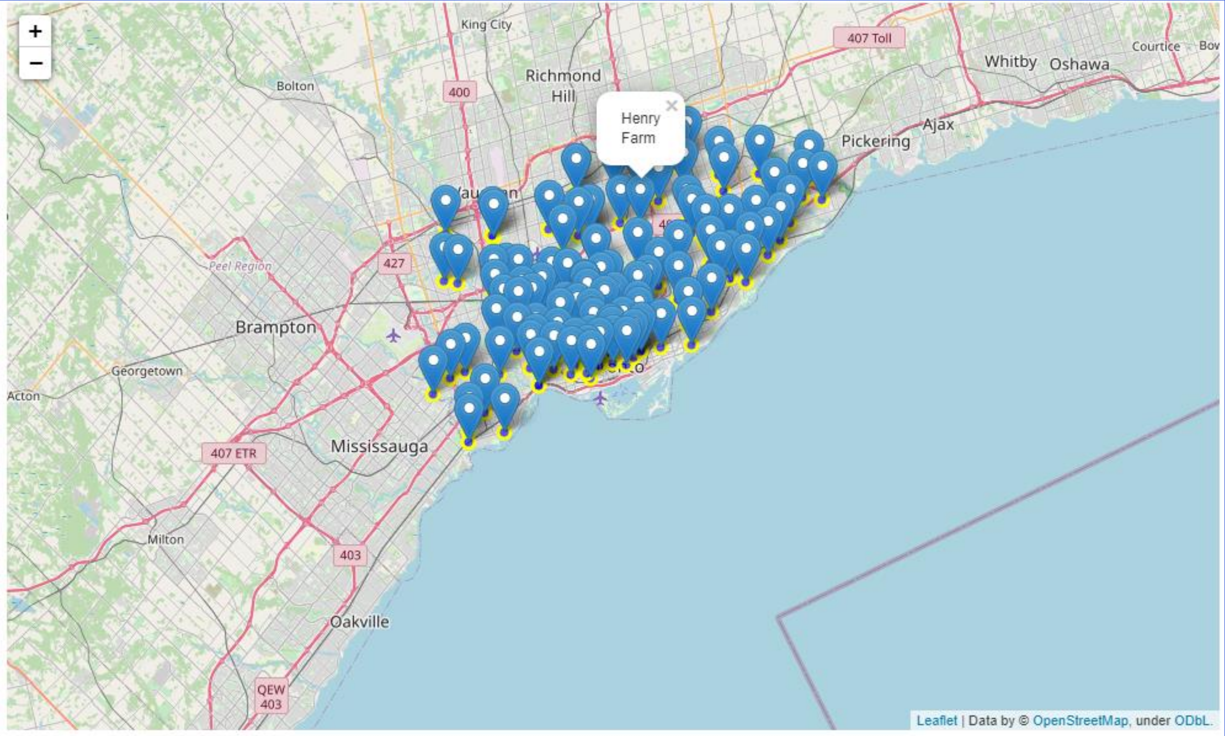
Ploting data with seaborn bar plot for each neighborhood that contain Crime Score



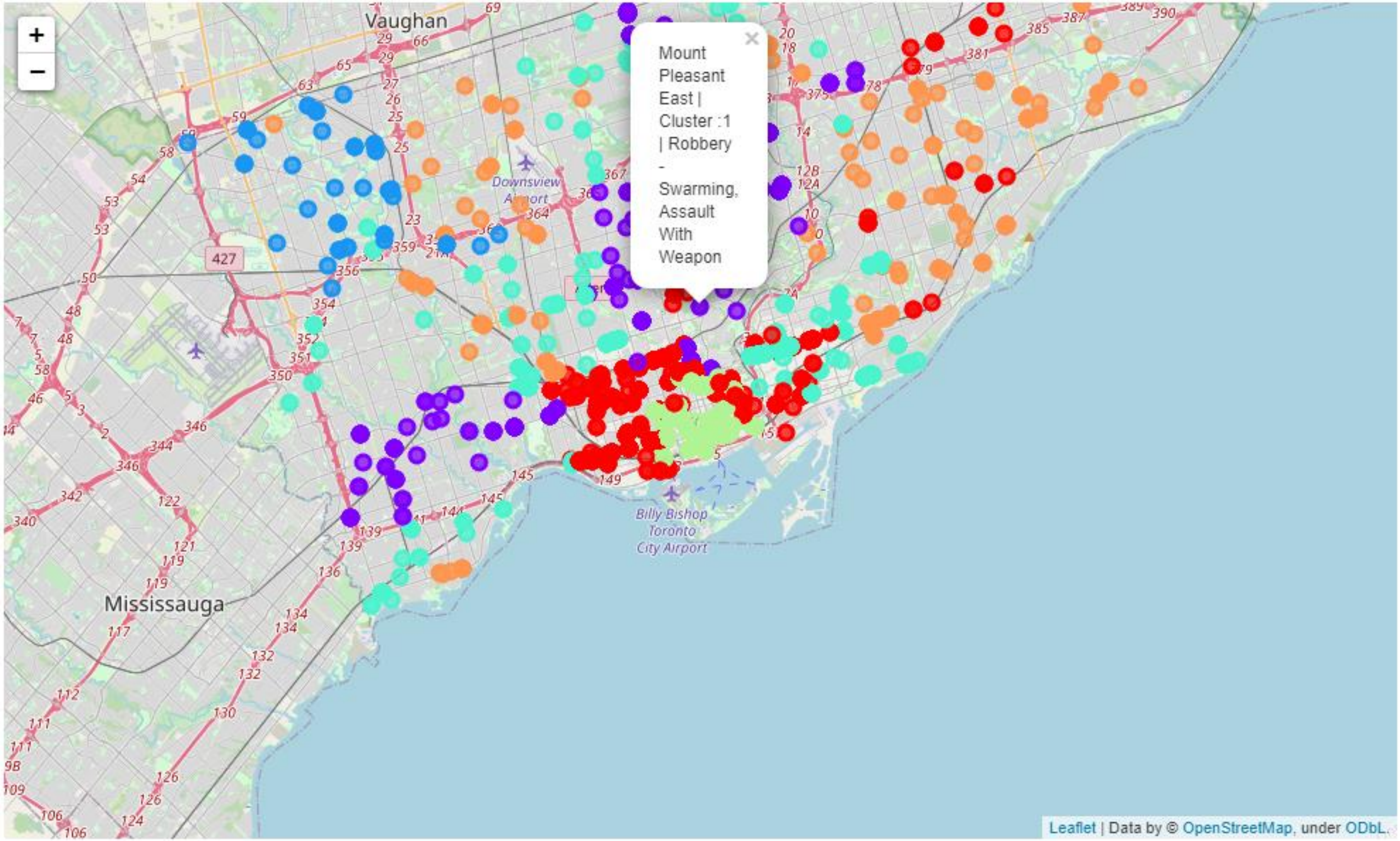
Map of Neighborhood

Visualizing Map of Neighborhood in Toronto with Folium. This map consist crime score for each Neighborhood in Toronto

Data obtained from Toronto Neighborhood Crime Score data and map created with use of Nominatim and Open Cage Geocode



Crime Incident Location by Cluster



Examine Venue Clusters based Crime Incidents

Example: Cluster 1

	Neighborhood	Venue	Venue Latitude	Venue Longitude	Venue Category	Crime Category	Cluster Labels	1st Most Common Crimes	2nd Most Common Crimes	3rd Most Common Crimes	4th Most Common Crime	5th Most Common Crime	6th Most Common Crime	7th Mos Commo Crim
0	Trinity Bellwoods	North of Brooklyn Pizzeria	43.646945	-79.406561	Pizza Place	Assault, B&E	0	Assault	B&E	Assault	B&E	Theft Of Motor Vehicle	Assault With Weapon	Assau Wit Weapo
1	Trinity Bellwoods	Soufi's	43.646553	-79.407339	Middle Eastern Restaurant	B&E, Assault, Theft Of Motor Vehicle, Theft Of...	0	Assault	B&E	Assault	B&E	Theft Of Motor Vehicle	Assault With Weapon	Assau Wit Weapo
2	Trinity Bellwoods	Frank & Oak	43.646579	-79.406655	Clothing Store	Theft Of Motor Vehicle	0	Assault	B&E	Assault	B&E	Theft Of Motor Vehicle	Assault With Weapon	Assau Wit Weapo
3	Trinity Bellwoods	Neat	43.646895	-79.405688	Furniture / Home Store	Robbery - Purse Snatch	0	Assault	B&E	Assault	B&E	Theft Of Motor Vehicle	Assault With Weapon	Assau Wit Weapo
4	Trinity Bellwoods	ZANE	43.646435	-79.407390	Accessories Store	Assault, B&E, Assault, Theft Of Motor Vehicle	0	Assault	B&E	Assault	B&E	Theft Of Motor Vehicle	Assault With Weapon	Assau Wit Weapo
...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
1946	Dovercourt Wallace Emerson Junction	Lynx Music	43.667425	-79.444994	Recording Studio	Assault	0	Assault	B&E	B&E	Assault	Theft Of Motor Vehicle	Assault With Weapon	B& W'Inter
1947	Dovercourt Wallace Emerson Junction	Jomar Electric Co. Ltd.	43.667360	-79.444795	Electronics Store	B&E, Assault, Theft Of Motor Vehicle,	0	Assault	B&E	B&E	Assault	Theft Of Motor Vehicle	Assault With Weapon	B& W'Inter

Conclusion Section

Project Conclusion

In this project, using k-means cluster algorithm I separated the crime incident into 10(Ten) different clusters and for 2150 different lattitude and logitude from dataset, which have very-similar crime incidents around them.

I feel rewarded with the efforts and believe this course with all the topics covered is well worthy of appreciation. This project has shown me a practical application to resolve a real situation that has impacting personal and safety impact using Data Science tools. The mapping with Folium is a very powerful technique to consolidate information and make the analysis and decision better with confidence.

Development Notes

This project can be continued for making it more precise in terms to analyze crime incident in Toronto for the next year (2020/2021).

Thanks for your attention