Exploring Neighborhoods of Toronto Using K-Means Clustering Algorithm and Toronto Crime Data

I. Introduction

1.1 Background

There are plenty of things one can gain from exploring different places. The list includes gaining new friends, new experiences, and new stories.

When you start exploring new places, you get a better understanding of the people living there, including their culture, history and background.

Studies show that travelling can improve your overall health and enhance your creativity. Therefore, you need to take time out from your daily tasks, office responsibilities, hectic schedule, and everyday pressures at least once in a year. Plan a tour to a new city with an open schedule and let life present you with the numerous opportunities. *[1]*

The planning and preparation stage for travel can be troublesome since it will take more time to generate a decent itinerary. As much as possible, we also want to explore unique places so that we can optimize travel time and cost. We may also include crime incidence in choosing the places we want to visit in a particular region.

1.2 Problem

As much as possible, we want to minimize the planning and the preparation stage for travel. Our safety is also included in identifying the places we want to visit. Thus, this project aims to demonstrate how to identify unique places to visit using foursquare and crime data.

1.3 Interest

This project would be beneficial to spontaneous travelers and backpackers who are interested in generating unique itineraries in a certain countries or places.

1.4 Scope and Limitation

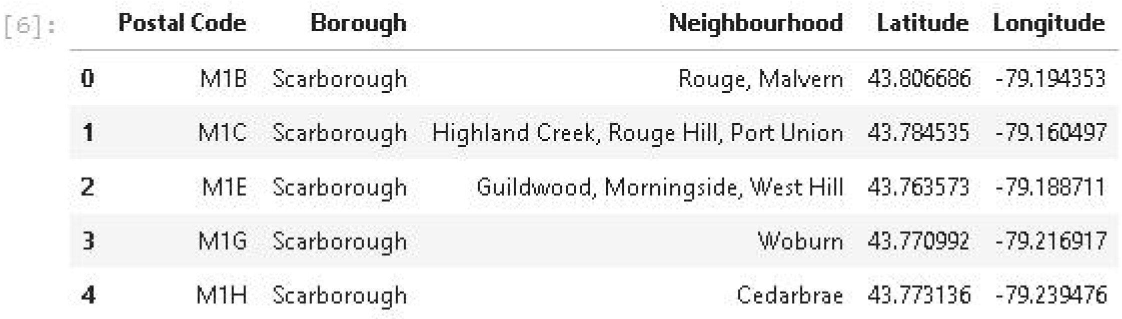
This project aims to segment and cluster the neighborhoods in East, West, and Central Toronto using foursquare data and will only be using the first 1,500 crime data from the Toronto Police data set. The K-Means algorithm will be used in clustering Toronto neighborhoods and will only be considering 5 clusters.

II. Data Sources, Acquisition and Cleaning

2.1 Postal Codes and Geographical Coordinates

Data source of postal code with corresponding boroughs and neighborhood were scraped and wrangled and cleaned from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. After the data cleaning for the postal code dataframe, it was combined with the dataframe from https://cocl.us/Geospatial_data to assign its corresponding geographic coordinates. Figure 1.1 is a sample of the final output.

Figure 1.1. Sample of Toronto Postal Code Dataset



[6]:

	Postal Code	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

2.1.1 Postal Codes Data Cleaning

To create the postal code dataframe from wikipedia, the following procedures were performed:

- Cells with assigned boroughs will only be processed and unassigned boroughs will be ignored.
- Postal codes with multiple assigned neighborhoods will be combined into one row with the assigned neighborhoods separated with a comma as shown in row 2 of Figure 1.1.
- Unassigned neighborhood of a borough will be assigned as the same as its corresponding borough name.

2.2 Neighborhood Venues, Toronto Major Crime Indicator (MCI) and MCI Data Cleaning

The Data source for neighborhood venue information was extracted from Foursquare API while the MCI was downloaded from Toronto Police Service Public Safety Data Portal. Rows with missing entries were deleted from the MCI dataframe.

III. Exploratory Data Analysis

3.1 Exploring and Mapping Toronto Neighborhoods

Figure 2.1 and 2.2 were created in order for us to locate the neighborhoods of Toronto and the neighborhoods we are interested in the analysis. Toronto has 11 boroughs and 103 neighborhoods in total and most of the neighborhoods are clustered near the harbour.

The East, west, and central Toronto has a total of 38 combined neighborhoods with a total of 835 venues and 191 unique categories.

Figure 2.1. Map of Toronto with Neighborhoods Superimposed on Top

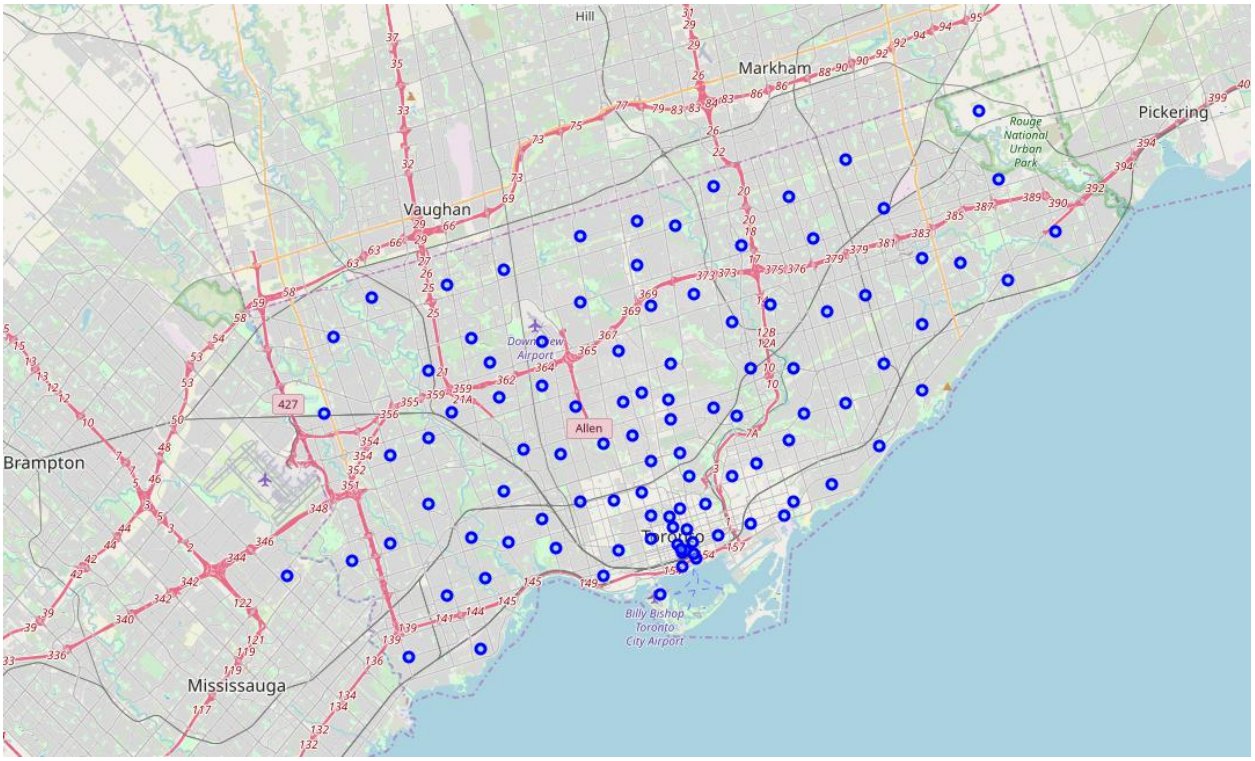
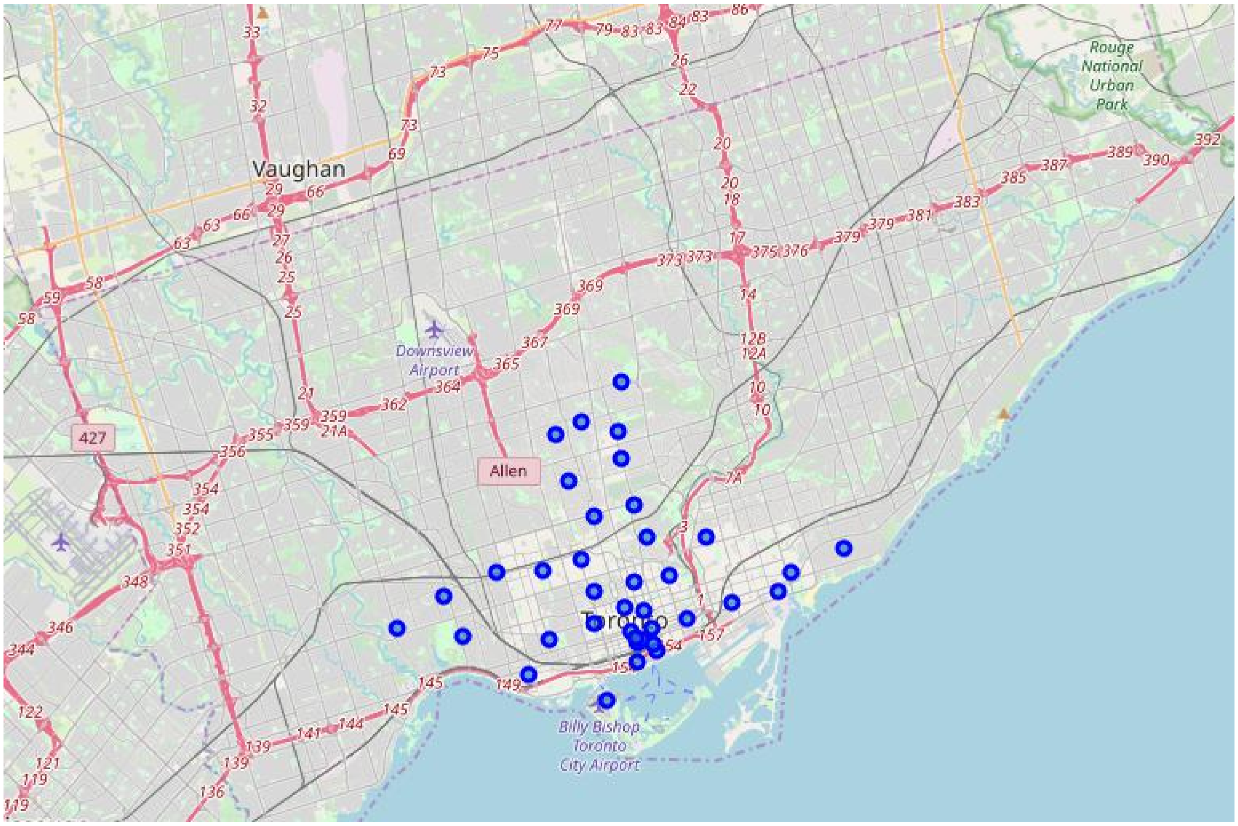


Figure 2.2. Map of East, West and Central Toronto with Neighborhoods Superimposed on Top



3.2 Mapping Toronto Major Crime Indicators

Figure 2.3 to 2.5 is the summary and mapping of Toronto's major crime indicator. Most of the crime activities is located in downtown of Toronto and most of the crime committed is assault. Therefore, we should be extra careful if we will be exploring downtown of Toronto.

Figure 2.3 Bar Chart of Toronto's Major Crime Indicator

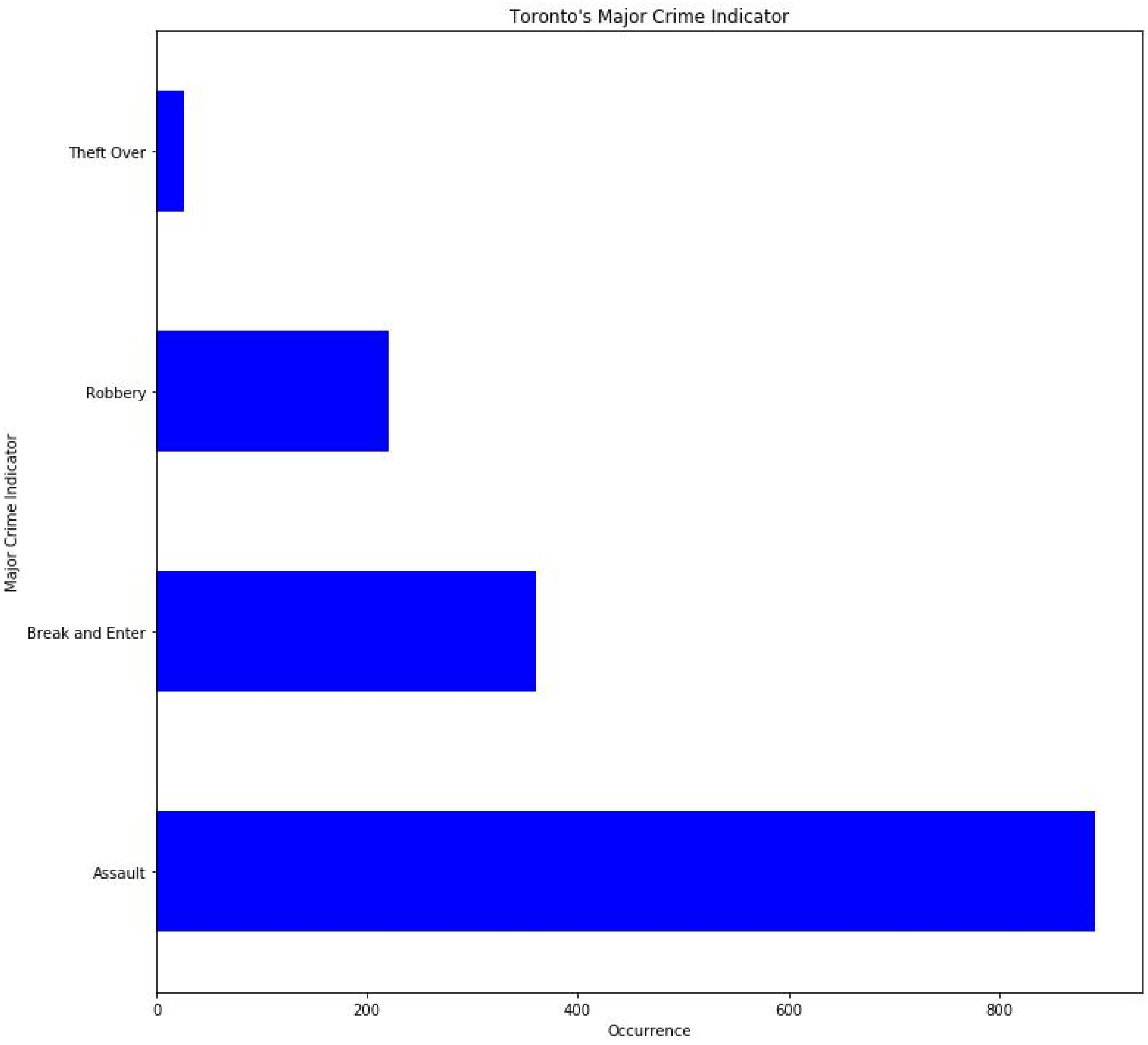


Figure 2.4 Scatter Plot of Toronto's Major Crime Indicator (*Sample Size = 1,500*)

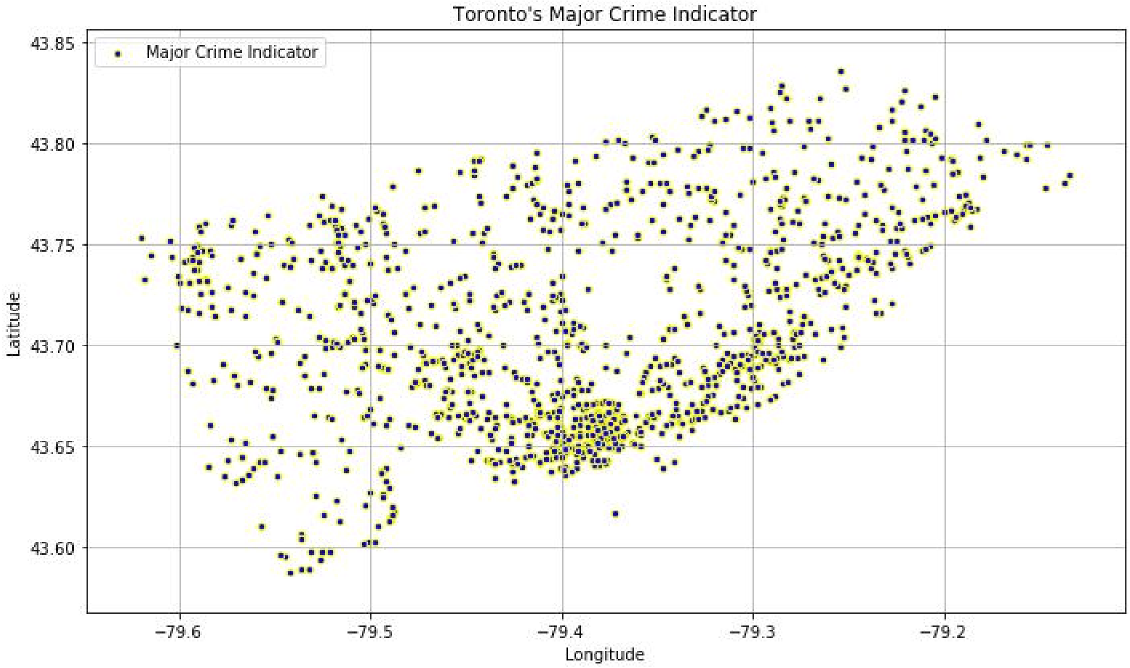
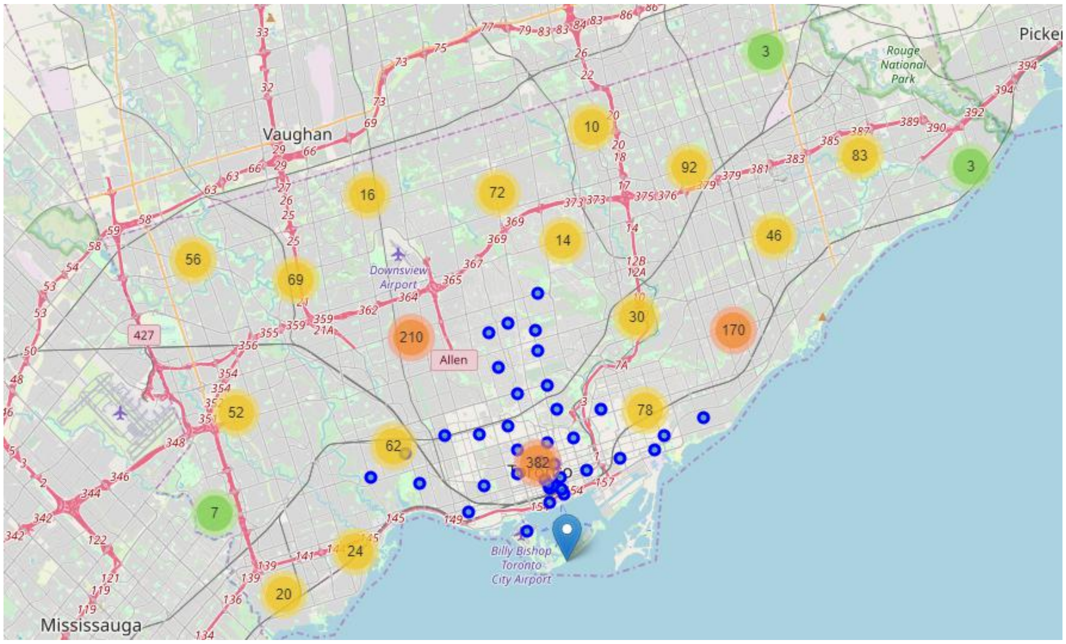


Figure 2.5 Folium Cluster Map of Toronto's Major Crime Indicator (*Sample Size = 1,500*)



IV. Clustering Neighborhoods

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. The main concept of the K-means algorithm is to represent each cluster by the vector of mean attribute values of all training instances for numeric attributes and by the vector of modal values for nominal attributes that are assigned to that cluster. This cluster representation is called cluster center.

In this project, we will cluster the neighborhoods of interest using K-Means clustering algorithm in selecting which neighborhoods to be included in the itinerary.

Figure 3.1 is the mapping of the clustered neighborhood using K-Means algorithm. Cluster Zero is widely scattered all over the neighborhood of interest as compared with the other cluster groups. This indicates that there are common venues all over the neighborhoods of interest and should avoid redundant visits to minimize travel costs and time.

Figure 3.2 is the mapping of major crime indicator and it shows that we should be extra careful in visiting neighborhoods near the harbour area. Table 1.1 is the proposed itinerary using K-Means algorithm when visiting east, west, and central Toronto including the top ten venues to visit in each neighborhood.

Figure 3.1 Map of East, West and Central Toronto of the Five Clustered Neighborhoods Superimposed on Top using K-Means Algorithm

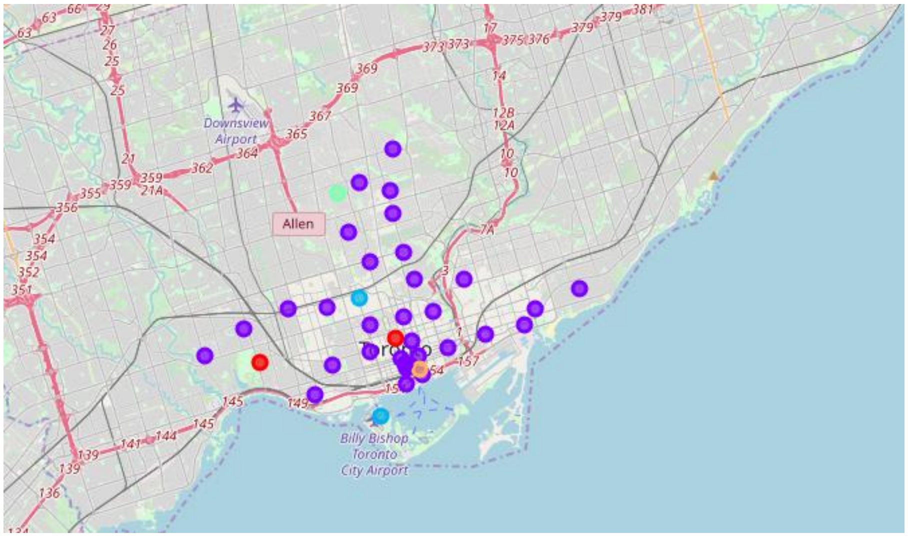


Figure 3.2 Map of East, West and Central Toronto of the Five Clustered Neighborhoods Superimposed on Top using K-Means Algorithm with Crime Data Location.

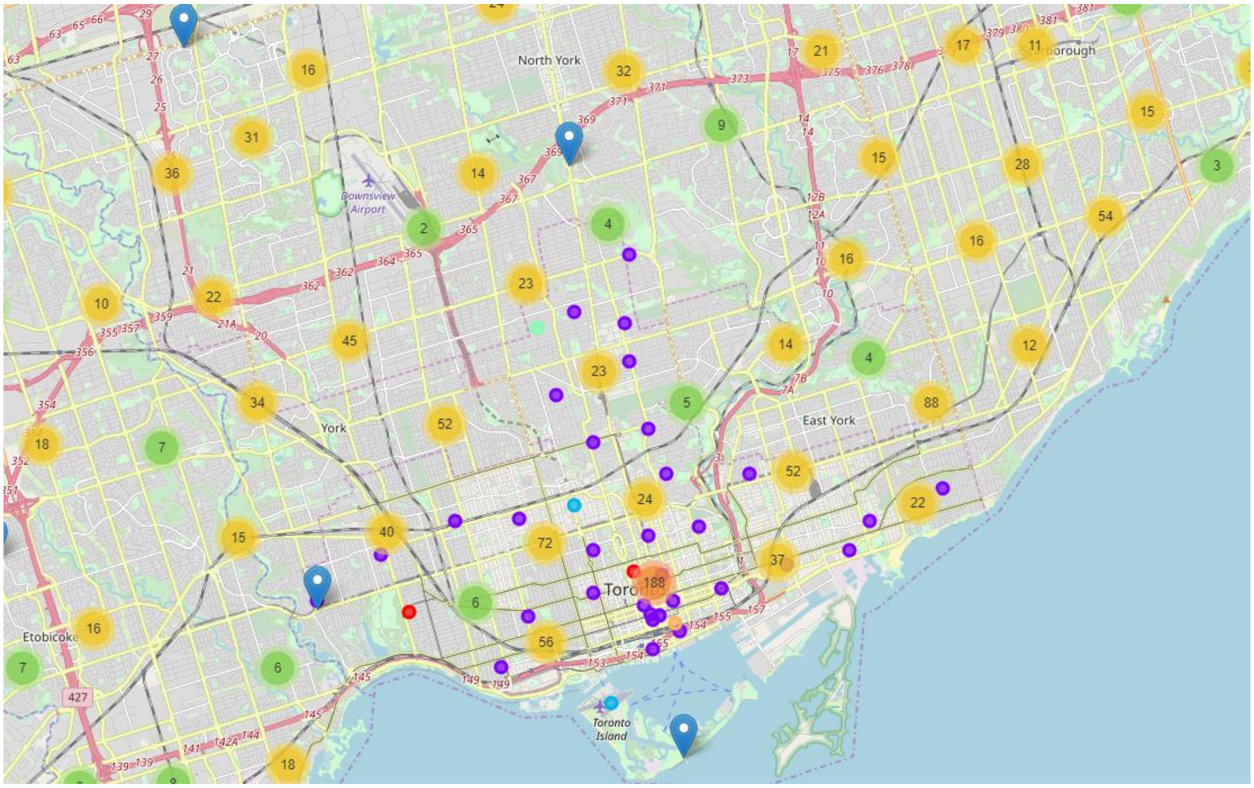


Table 1.1 Summary of Proposed Toronto Itinerary base from K-Means Clustering Algorithm and Toronto Major Crime Indicator.

Itinerary	Cluster Group	Borough	Neighbourhood	Top Venue Ranking	Venue
Day 1 - 3	1	Downtown Toronto	Stn A PO Boxes 25 The Esplanade *(High Criminal Activity)*	1	Farmers Market
				2	Cocktail Bar
				3	CafÃ©
				4	Seafood Restaurant
				5	Beer Bar
				6	Food Truck
				7	Museum
				8	Clothing Store
				9	Pub
				10	Jazz Club
Day 4 - 6	4	Downtown Toronto	Central Bay Street *(High Criminal Activity)*	1	Coffee Shop
				2	CafÃ©
				3	Chinese Restaurant
				4	Italian Restaurant
				5	Bubble Tea Shop
				6	Sandwich Place
				7	Ice Cream Shop
				8	Spa
				9	Japanese Restaurant
				10	Seafood Restaurant

Day 7 - 9	3	Central Toronto	The Annex, North Midtown, Yorkville *(Low Criminal Activity)*	1	Sandwich Place
				2	CafÃ©
				3	Coffee Shop
				4	Burger Joint
				5	Pizza Place
				6	Pub
				7	Middle Eastern Restaurant
				8	BBQ Joint
				9	History Museum
				10	Liquor Store
Day 10 - 12	0	Central Toronto	Forest Hill North, Forest Hill West *(Low Criminal Activity)*	1	Jewelry Store
				2	Trail
				3	Sushi Restaurant
				4	Bus Line
				5	Yoga Studio
				6	Diner
				7	Event Space
				8	Ethiopian Restaurant
				9	Eastern European Restaurant
				10	Dumpling Restaurant
Day 15 to 15	2	Central Toronto	Roselawn *(Low Criminal Activity)*	1	Garden
				2	Yoga Studio
				3	Dim Sum Restaurant

				4	Falafel Restaurant
				5	Event Space
				6	Ethiopian Restaurant
				7	Eastern European Restaurant
				8	Dumpling Restaurant
				9	Donut Shop
				10	Dog Run

V. Conclusion and Recommendation

In this study, we were able to generate a proposed itinerary using K-Means algorithm. The identified neighborhoods are one of the possible unique neighborhoods to visit. The major criteria in generating Table 1.1 is the proximity of the unique neighborhoods in order to reduce travel time and costs and to maximize the length of stay of the proposed neighborhoods to visit. Crime data was also incorporated in the map so that we will be aware of our safety when exploring the neighborhoods to explore.

This method would be beneficial to those travelers who are interested in planning their itinerary that can can minimize travel costs and time and can maximize exploring unique neighborhoods by identifying similar and unique neighborhoods.

Please take note that this study focuses only with the Foursquare API data in clustering the neighborhoods using K-Means algorithm. It is suggested to also include the major crime indicator data in clustering the neighborhoods for better results.