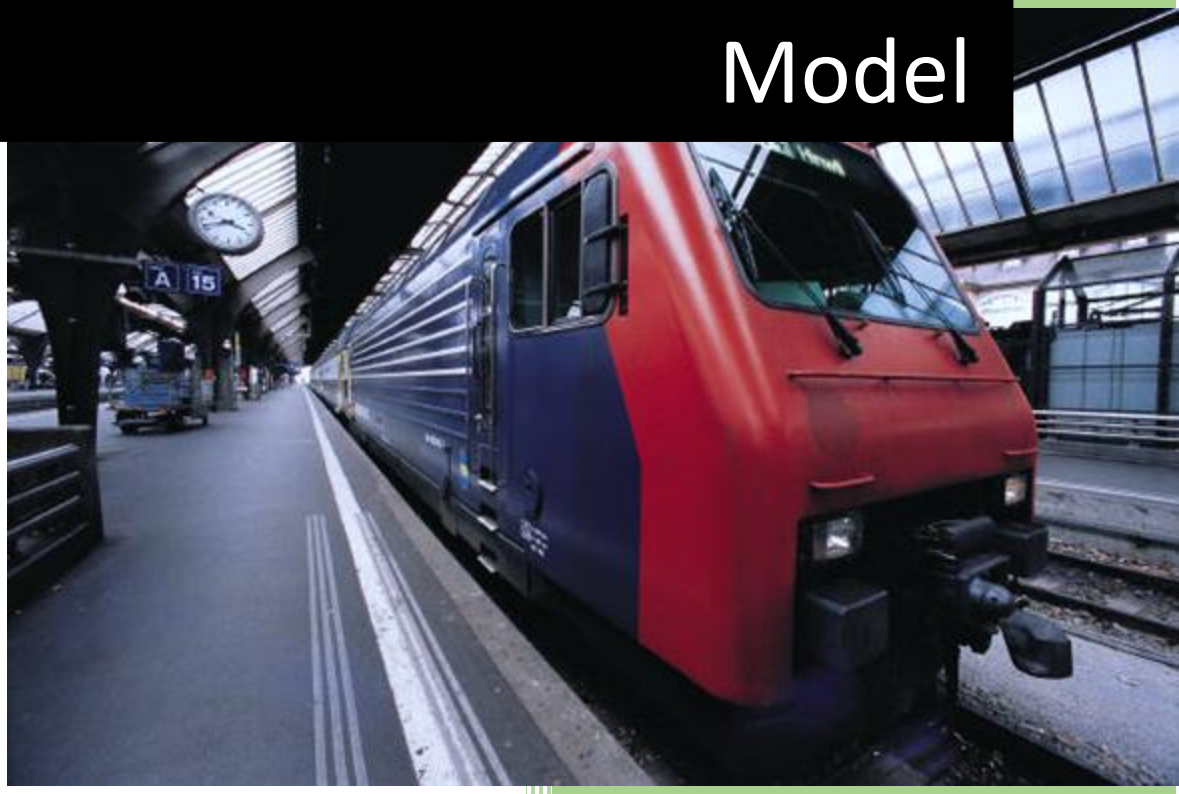
Neighborhood Recommendation Model



Bitun Sen

4/5/2020

CONTENTS

INTRODUCTION

In the era of globalization, people moves a lot looking for better life and better work opportunities. Sometimes, this move becomes very hectic when it is to a completely unknown country or place. We people always have our own preference which might not match with other person. For example, my brother. He lives in India and works for a major US based IT firm. His company got a very good project to be executed for a client based out of Toronto, Canada. So my brother's company wants him to move to Toronto, Canada.

This is the first time, my brother is moving outside of India and is completely confused in selecting the neighborhood for his staying. After doing almost a month's research, he reached out to me to help him out in selecting or recommending the neighborhood which will be best for him to stay.

When I started enquiring about his preferences, he came up with the following to be used as selection criteria:

- Safe Neighborhood
- Transportation - Metro Station, Bus Station
- Breakfast places
- Grocery Shops
- Coffee Shops
- Restaurants (Indian, Italian, Thai and American)
- Shopping Center
- Selected Bars - Sports Bar, Cocktail Bar, Pub
- Outdoor Activity Center - Playground, Park

He also had provided his office address at Toronto. He wants the preferred neighborhoods should be within 6 miles from his office address.

With these criteria provided by my brother, I would like to recommend to my brother which neighborhood of Toronto, Canada will be good for his living. The basis of this work will help the people like my brother to find their kind of neighborhood considering their preferences in mind.

DATA ACQUISITION

To implement a recommendation system for Toronto neighborhoods, I will look for the below data sources:

- Neighborhood data source
- Crime data source
- Preferred Venue data using Foursquare API

NEIGHBORHOOD DATA SOURCE

In Toronto, there are total 140 neighborhoods. To get the list of these neighborhood and their boundary information, I will be using the data from the below URL:

http://data.torontopolice.on.ca/datasets/neighbourhood-crime-rates-boundary-file-

The data available in this page holds the below information:

- Name of the neighborhood
- Geometric information about the neighborhood (coordinates of the boundary)
- Average crime rate for specific crimes till 2018.

I will use an Open Data Portal based API to access this data. The API URL is:

https://opendata.arcgis.com/datasets/af500b5abb7240399853b35a2362d0c0_0.geojson

From this dataset, I will use only the name of neighborhood, the boundary coordinates. Crime data will be skipped due to old information (data till 2018). This dataset has the boundary coordinates of each neighborhood, which is not suitable for getting the venues. As, the boundary coordinates are nothing but the coordinates of a polygon, that is why using the boundary coordinates, I have calculated the centroids of all the neighborhood and assigned them as centroid coordinates of each neighborhood.

Python's geopy package will be used to retrieve the coordinates of the office address. Then for each neighborhood, the distance from the office will be calculated.

CRIME DATA

Basically Toronto is one of the safest place in the world, but in recent years, crime has started rising in different neighborhoods. We will use the crime incident data published by Toronto Police Department. This data set captured all the major crime reports since 2014 covering for all the 140 neighborhoods. The URL is: http://data.torontopolice.on.ca/datasets/mci-2014-to-2019.

I have used the API exposed by Open Data Portal to access this data programmatically. The service URL is:

https://opendata.arcgis.com/datasets/f4c2e5de021f4836a3caf77f8421f487_0.geojson.

This dataset also contains the coordinates of all the places of the incidents. These coordinate information is not relevant here, so I will ignore those information.

In this dataset, the name of the neighborhood was having some additional information (HOOD ID, some number related to Toronto police data), which I had to strip it out so that both the neighborhood dataset and crime dataset can have same neighborhood name.

As safety is his one of the priority, rather than considering all the 140 neighborhoods, I will take first 100 lowest crime based neighborhood.

FOURSQUARE VENUE DATA

My brother has shared his preferred venues which he wants near to his future place of residence. It is quite possible that all venues might not be available at all neighborhood. To collect the list of preferred

venues, I will be using the category hierarchy list to identify the category Identifier of each preferred categories to filter out the non-preferred venues. The link for the Foursquare category list is below:

https://developer.foursquare.com/docs/build-with-foursquare/categories/

Going through this page, at first I will collect all the nearest and closest categoryIds for all the preferred facilities my brother is looking for. Then I will use the Foursquare API for searching the venues by passing all these categoryIds.

We will use clustering technique to group the neighborhoods based on their availability near to each neighborhood and then come up with the recommendations.

METHODOLOGY

My brother emphasizes on public safety while choosing the preferred neighborhoods. So, I will at first collect the data which will help me to find the safest neighborhoods.

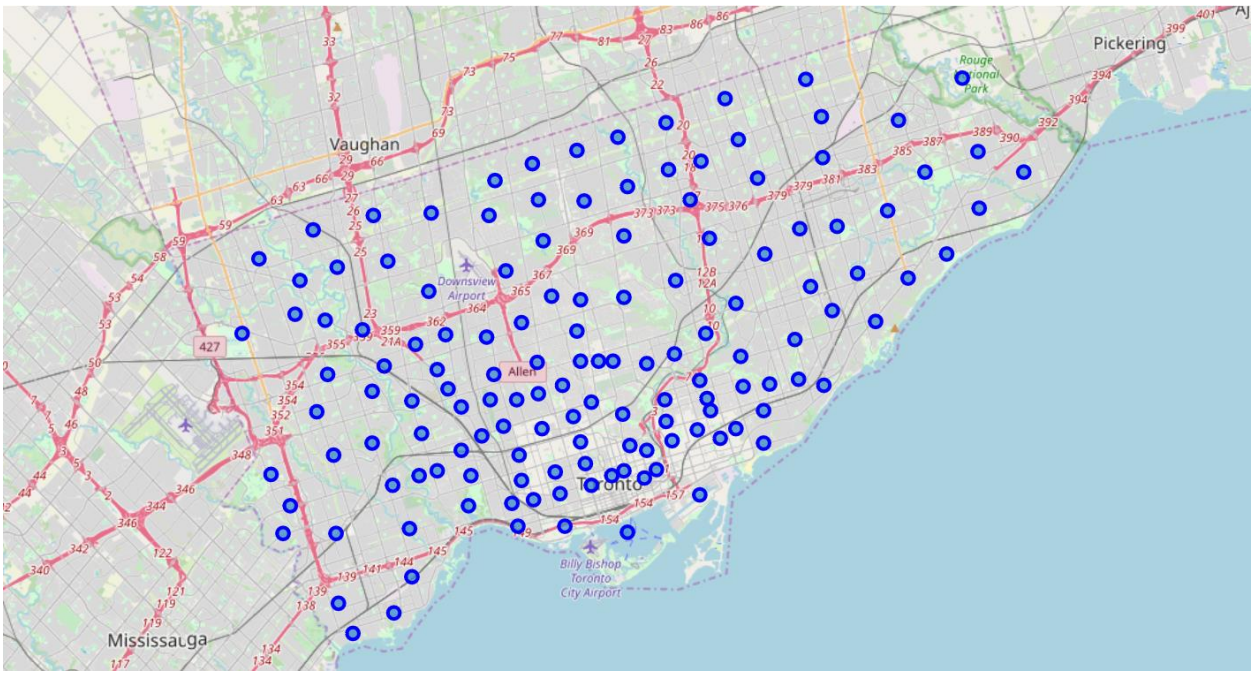
I will use Python as the programming language and its different packages which will be relevant for this whole exercise.

COLLECTING BASIC NEIGHBORHOOD INFORMATION

Using the neighborhood dataset provided by the Toronto Department of Police, I have retrieved the name of the neighborhood and then calculated the centroid (coordinates) of each neighborhood using the boundary coordinates present in the dataset.

	Neighborhood	Latitude	Longitude	Distance_From_Office
0	Yonge-St.Clair	43.687859	-79.397831	1.36
1	York University Heights	43.765738	-79.488842	8.40
2	Lansing-Westgate	43.754272	-79.424706	6.10
3	Yorkdale-Glen Park	43.714673	-79.457068	4.76
4	Stonegate-Queensway	43.635520	-79.501091	6.39
5	Tam O'Shanter-Sullivan	43.780130	-79.302876	8.54
6	The Beaches	43.671049	-79.299560	4.20
7	Thistletown-Beaumond Heights	43.737989	-79.563452	10.13
8	Thorncliffe Park	43.707749	-79.349944	3.03
9	Danforth East York	43.689468	-79.331362	2.90

Using Python's folium package, I plotted all the neighborhoods on Toronto Map.



COLLECTING CRIME DETAILS FOR TORONTO NEIGHBORHOODS

Using the major criminal incident dataset published by Toronto Department of Police, I have retrieved all the major criminal incidents per neighborhood which happened since 2014 till 2019. Then I have calculated total number of major criminal incidents per neighborhood. Then I have calculated the percentage of criminal incidents per neighborhood by dividing the total number of criminal incidents in each neighborhood by the total number incidents in Toronto. Then I sorted the dataset by the percentage of criminal incidents in ascending order.

	Neighborhood	Crime_Count	Percentage_Crime
0	Lambton Baby Point	353	0.001710
1	Woodbine-Lumsden	377	0.001826
2	Maple Leaf	410	0.001986
3	Guildwood	411	0.001991
4	Yonge-St.Clair	412	0.001996
5	Markland Wood	413	0.002001
6	Old East York	479	0.002320
7	Casa Loma	480	0.002325
8	Forest Hill South	494	0.002393
9	Kingsway South	496	0.002403

This tells us that “Lambton Baby Point” neighborhood is the safest neighborhood based on the criminal incidents published publicly. If we see that last 10 records of the dataset, that will tell us which neighborhoods had higher crime reported.

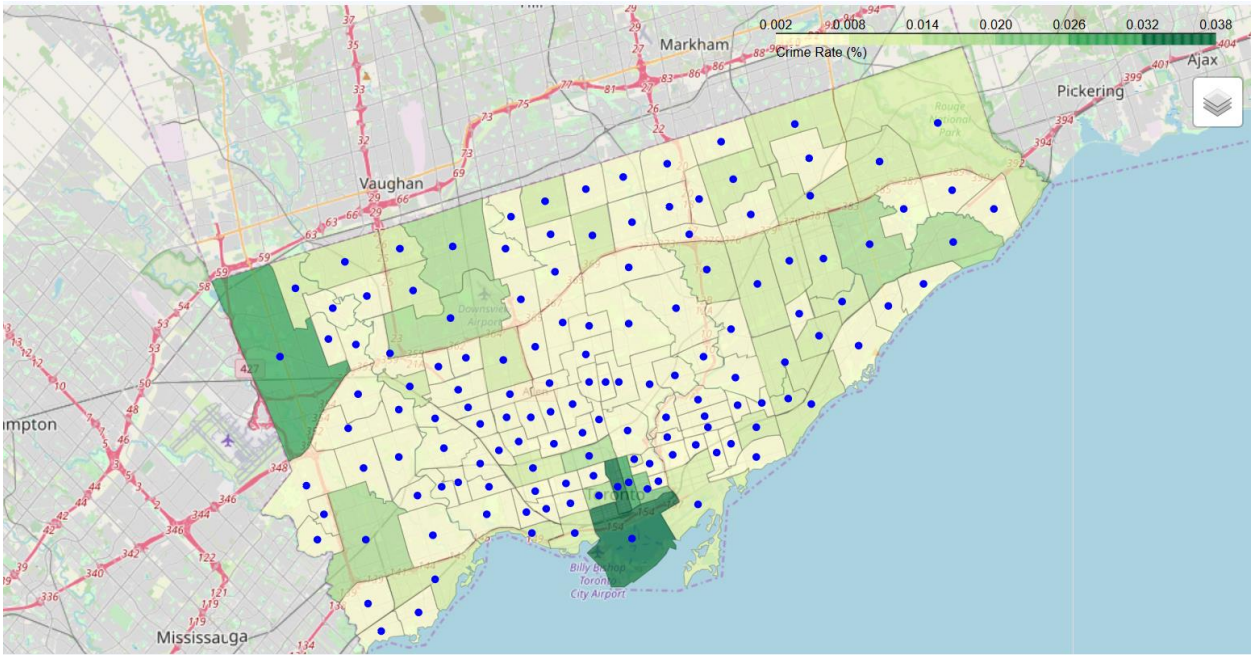
	Neighborhood	Crime_Count	Percentage_Crime
130	West Hill	3497	0.016940
131	Woburn	3798	0.018398
132	Kensington-Chinatown	3823	0.018519
133	Downsview-Roding-CFB	3974	0.019251
134	York University Heights	3989	0.019323
135	Moss Park	4786	0.023184
136	West Humber-Clairville	5702	0.027621
137	Church-Yonge Corridor	6232	0.030189
138	Bay Street Corridor	6817	0.033023
139	Waterfront Communities-The Island	7747	0.037528

So the highest crime rate is at Waterfront Communities-The Island. Compare to the number of incidents at Lambton Baby Point, the number of incidents at Waterfront communities is quite at higher side.

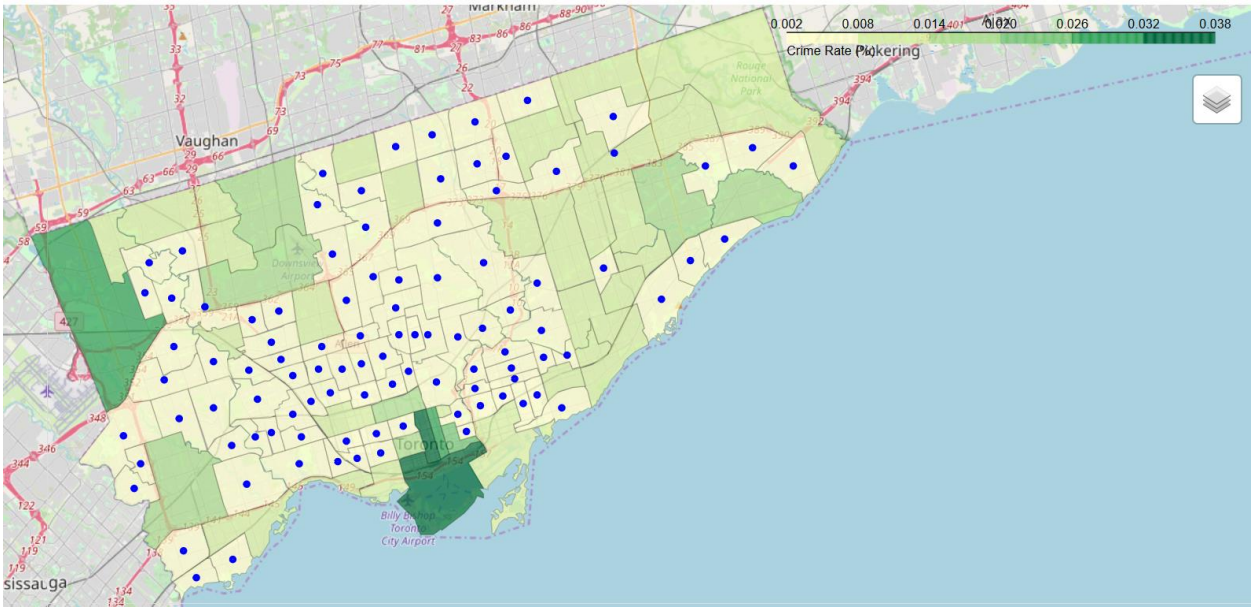
Then I have merged both the dataset using the name of the neighborhood to have a consolidated dataset for each neighborhood.

	Neighborhood	Crime_Count	Percentage_Crime	Latitude	Longitude	Distance_From_Office
0	Lambton Baby Point	353	0.001710	43.657421	-79.496008	5.72
1	Woodbine-Lumsden	377	0.001826	43.694107	-79.311123	3.96
2	Maple Leaf	410	0.001986	43.715575	-79.480718	5.76
3	Guildwood	411	0.001991	43.748827	-79.195014	10.85
4	Yonge-St.Clair	412	0.001996	43.687859	-79.397831	1.36
5	Markland Wood	413	0.002001	43.633542	-79.573394	9.87
6	Old East York	479	0.002320	43.696781	-79.335448	2.99
7	Casa Loma	480	0.002325	43.681852	-79.407967	1.43
8	Forest Hill South	494	0.002393	43.694526	-79.414278	2.23
9	Kingsway South	496	0.002403	43.653522	-79.510540	6.48

Then I have created one map using folium's Choropleth utility to visualize the neighborhoods based on their criminal incident rates. The darker shade denotes the higher criminal incident rates.

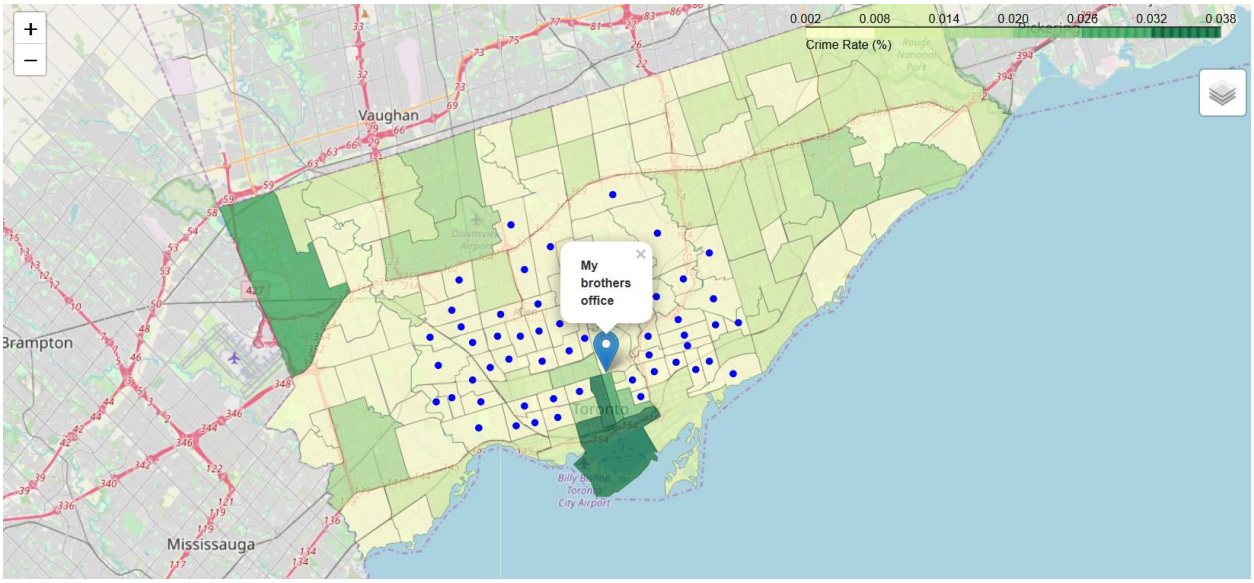


Being first time away from his home country and moving to an unknown place, my brother was preferring safer neighborhood. That is why, I have decided to continue my next part of analysis using the first 100 neighborhoods.



FILTER NEIGHBORHOOD BY DISTANCE FROM OFFICE

Now, I have applied another filter criteria, distance of neighborhood from the office. After applying that “6 miles” filter, I got total 92 neighborhoods left for further analysis.



RETRIEVE LIST OF PREFERRED VENUES

My brother shared a list of his preferred venues which he feels to be good to have near his neighborhood. The list is below:

- Safe Neighborhood
- Transportation - Metro Station, Bus Station
- Breakfast places
- Grocery Shops
- Coffee Shops
- Restaurants (Indian, Italian, Thai and American)
- Shopping Center
- Selected Bars - Sports Bar, Cocktail Bar, Pub
- Outdoor Activity Center - Playground, Park

I have used the category list, provided by Foursquare, to find out the list of category Ids. The link is provided below:

- Foursquare category hierarchy

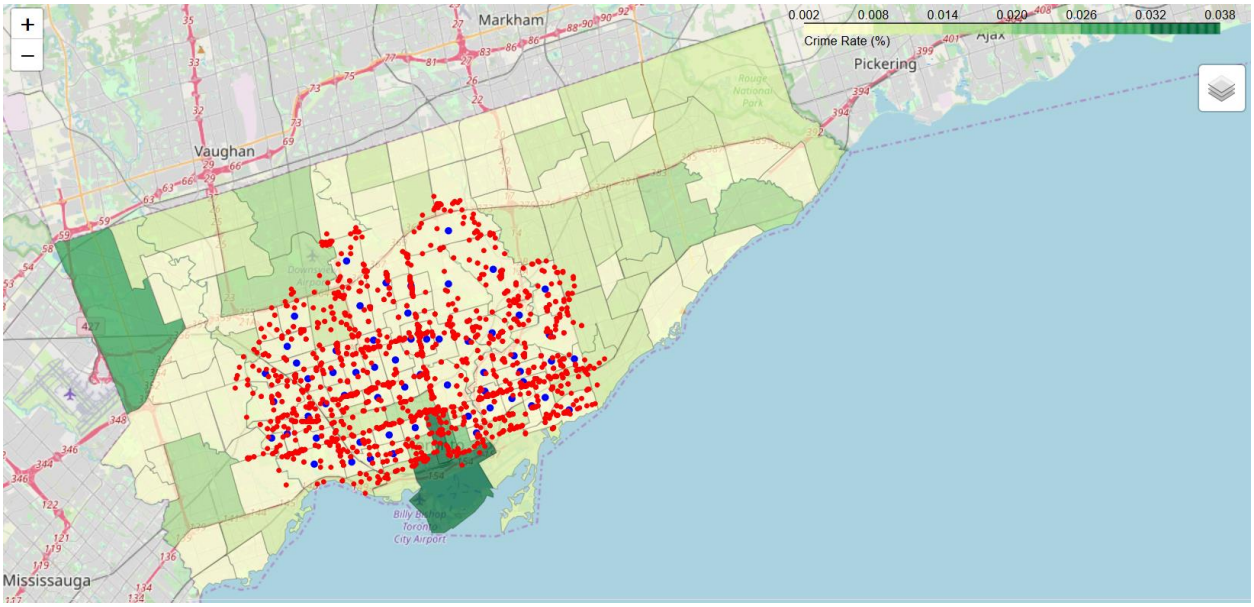
To meet my brother's venue preference, I had ended up getting total 17 unique categories:

```
CATEGORY_ID_QUERY = ("4bf58dd8d48988d1fd931735,4bf58dd8d48988d10f941735,4bf58dd8d48988d110941735,4bf58dd8d48988d149941735"
                     ",4bf58dd8d48988d14e941735,4bf58dd8d48988d1fd941735,5744ccdfe4b0c0459246b4dc,4bf58dd8d48988d11d941735"
                     ",4bf58dd8d48988d11e941735,4bf58dd8d48988d11b941735,4bf58dd8d48988d1e7941735,4bf58dd8d48988d163941735"
                     ",4bf58dd8d48988d143941735,52e81612bcbc57f1066b79f4,4bf58dd8d48988d1e0931735,52f2ab2ebcbc57f1066b8b4f"
                     ",52f2ab2ebcbc57f1066b8b42")
```

Foursquare search venue API has been leveraged to find out the preferred venues by providing the neighborhood's coordinates and above mentioned category Ids of the preferred venues. The search of venues were limited with 1500 meter radius from the coordinate of the neighborhoods.

Using 59 neighborhoods, I created a dataset which consists of the below information:

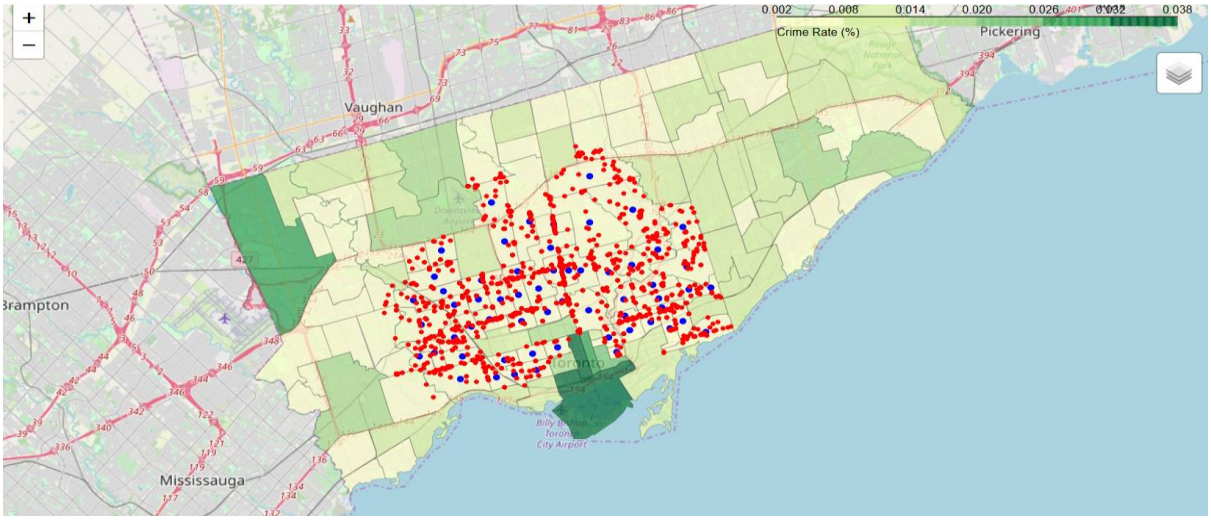
- Name of the neighborhood
- Name of the Venue
- Name of Venue Category
- Coordinates (latitude, longitude) of the venue



In this above map, the red circles are showing the venues and the blue dots are showing the centroids of the filtered neighborhoods.

From the above map, I could see that there are some venues which are inside the neighborhoods which we had discarded before due to higher criminal activities. Let's discard those venues which are in those unsafe neighborhoods.

To check if a venue is within the boundary of unsafe neighborhood, I will use Python's shapely package. After removing those venues, the map of selected venues look like:



With this process, the number of preferred venues went down from 2542 to 2032. Now let's transform this dataset so that it can be used for clustering.

For better understanding of the number of preferred venues per neighborhood, I have created a separate dataset which is sorted in descending order on number of venues (Category column in below picture denotes the number of venues):

	Neighborhood	Category
0	Humewood-Cedarvale	49
1	Forest Hill South	48
2	St.Andrew-Windfields	48
3	Oakwood Village	48
4	Caledonia-Fairbank	47
5	Keelesdale-Eglinton West	47
6	Lambton Baby Point	46
7	Flemingdon Park	46
8	Rockcliffe-Smythe	46
9	Leaside-Bennington	45

My plan is to create cluster of neighborhoods based on the number of venues present in each neighborhood. There are total 17 unique categories which are categorical features. Using one hot encoding, I transformed the dataset. After adding the name of the neighborhood in this new dataset, now the dataset has total 18 columns:

	Neighborhood	American Restaurant	Big Box Store	Breakfast Spot	Buffet	Bus Stop	Cocktail Bar	Coffee Shop	Indian Restaurant	Italian Restaurant	Metro Station	Park	Playground	Pub	Shopping Mall	Shopping Plaza	Sports Bar	Thai Restaurant
0	Lambton Baby Point	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1	Lambton Baby Point	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
2	Lambton Baby Point	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Lambton Baby Point	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
4	Lambton Baby Point	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0

After this I have converted this dataset into the representation by neighborhood. For this I had applied groupby function and then applied mean for each venue category. That transformed the dataset as below:

	Neighborhood	American Restaurant	Big Box Store	Breakfast Spot	Buffet	Bus Stop	Cocktail Bar	Coffee Shop	Indian Restaurant	Italian Restaurant	Metro Station	Park	Playground	Pub	Shopping Mall	Shopping Plaza	Sports Bar	Tha Restaurar
0	Banbury-Don Mills	0.028571	0.000000	0.028571	0.0	0.028571	0.000000	0.257143	0.057143	0.142857	0.000000	0.314286	0.000000	0.000000	0.057143	0.000000	0.000000	0.08571
1	Bedford Park-Nortown	0.066667	0.000000	0.000000	0.0	0.066667	0.000000	0.222222	0.111111	0.200000	0.000000	0.177778	0.000000	0.044444	0.022222	0.000000	0.000000	0.08888
2	Beechborough-Greenbrook	0.088889	0.022222	0.022222	0.0	0.066667	0.022222	0.200000	0.000000	0.044444	0.022222	0.333333	0.044444	0.022222	0.066667	0.000000	0.044444	0.00000
3	Blake-Jones	0.074074	0.000000	0.000000	0.0	0.000000	0.037037	0.222222	0.111111	0.037037	0.185185	0.148148	0.000000	0.148148	0.000000	0.037037	0.000000	0.00000
4	Briar Hill-Belgravia	0.029412	0.000000	0.029412	0.0	0.088235	0.029412	0.117647	0.058824	0.117647	0.029412	0.294118	0.029412	0.000000	0.088235	0.000000	0.029412	0.05882

Now this dataset I will use for K-mean clustering.

RESULTS

For K-mean clustering, we need to predefine the number of clusters and provide that number as input while doing the K-mean clustering. I had used 3, 4 and 5 as number of clusters. After analyzing the created clusters, I felt like 3 is the correct and meaningful number for this dataset.

After assigning the cluster number to the respective neighborhood, I have sorted the venue categories based on the number of venues per category for each neighborhood. This gives a clear information, which venue category is more popular in each neighborhood. For example: Coffee Shop is the most preferred venue at Lambton Baby Point neighborhood.

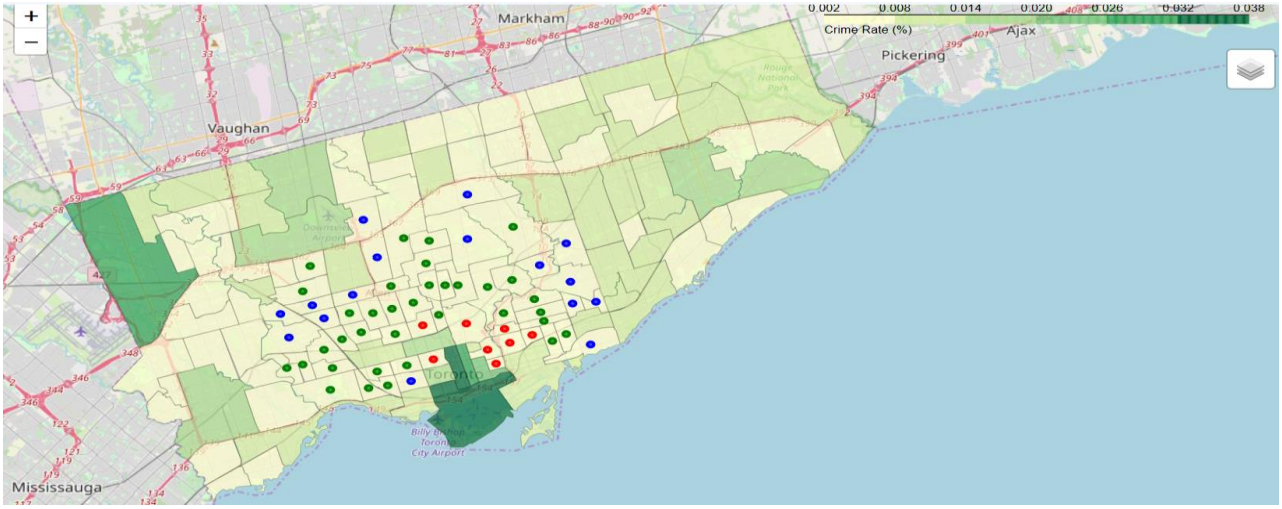
If any venue is not present in any particular neighborhood, I have added "NA" after the venue category. For example: Lambton Baby Point does not have any Sports Bar or Buffet or Shopping Plaza. That is why in the last 4 columns for this neighborhood, those columns have:

- Sports Bar NA
- Shopping Mall NA
- Buffet NA
- Shopping Plaza NA

	Neighborhood	Crime_Count	Distance_From_Office	Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	...	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue	16th Most Common Venue	17th Mo Commo Venu
0	Lambton Baby Point	353	5.72	46	2	Coffee Shop	Park	Italian Restaurant	Playground	Metro Station	...	Big Box Store	American Restaurant	Sports Bar NA	Shopping Mall NA	Buffet NA	Shoppir Plaza N
1	Woodbine-Lumsden	377	3.96	31	1	Bus Stop	Coffee Shop	Park	Indian Restaurant	Breakfast Spot	...	Pub NA	Shopping Mall NA	Buffet NA	Shopping Plaza NA	Big Box Store NA	America Restaura N
2	Maple Leaf	410	5.76	22	2	Coffee Shop	Park	American Restaurant	Bus Stop	Playground	...	Indian Restaurant NA	Metro Station NA	Cocktail Bar NA	Pub NA	Buffet NA	Shoppir Plaza N
3	Yonge-St.Clair	412	1.36	39	2	Coffee Shop	Park	Pub	Metro Station	Italian Restaurant	...	Cocktail Bar NA	Bus Stop NA	Buffet NA	Shopping Plaza NA	Big Box Store NA	Th Restaura N
4	Old East York	479	2.99	45	2	Coffee Shop	Park	Indian Restaurant	Thai Restaurant	Bus Stop	...	Playground NA	Sports Bar NA	Cocktail Bar NA	Shopping Plaza NA	Buffet NA	Big Bo Store N

I have plotted the neighborhoods in the map as well for better visualization. In this map, below are the legends:

- RED denotes Cluster 0
- BLUE denotes Cluster 1
- GREEN denotes Cluster 2



DISCUSSION

From the clusters plotted on the map is giving a very good representation about neighborhoods. The neighborhoods of cluster 0 (RED in color) are having less preferred venues. The reason being, most of the venues got filtered out due to their geographical location falls within unsafe neighborhoods.

Then comes Cluster 1 (BLUE in color). These neighborhoods are also adjacent to other unsafe neighborhoods which forces the preferred venues to ignore.

The Cluster 2 neighborhoods (GREEN in color) are the safest neighborhoods and also surrounded by safe neighborhoods. That is why those are having more preferred venues. As Toronto is one of the safest city in the world, obviously this cluster has more members compare to other clusters.

So considering all the input criteria my brother had given to me, I would recommend couple of neighborhoods from Cluster 2. From safety point of view, though it is far from his office, I would recommend “Lambton Baby Point”, because it is having the lowest criminal incidents reported. It also has almost all venues my brother had mentioned.

Maple Leaf will be missed as it doesn’t have one single Indian restaurant.

I would also like to recommend “Old East York” which has “Indian Restaurant as the third popular venue and also it is closer to his office.

	0
Neighborhood	Lambton Baby Point
Crime_Count	353
Distance_From_Office	5.72
Category	46
Cluster Labels	2
1st Most Common Venue	Coffee Shop
2nd Most Common Venue	Park
3rd Most Common Venue	Italian Restaurant
4th Most Common Venue	Playground
5th Most Common Venue	Metro Station
6th Most Common Venue	Breakfast Spot
7th Most Common Venue	Bus Stop
8th Most Common Venue	Thai Restaurant
9th Most Common Venue	Pub
10th Most Common Venue	Indian Restaurant
11th Most Common Venue	Cocktail Bar
12th Most Common Venue	Big Box Store
13th Most Common Venue	American Restaurant
14th Most Common Venue	Sports Bar NA
15th Most Common Venue	Shopping Mall NA
16th Most Common Venue	Buffet NA
17th Most Common Venue	Shopping Plaza NA

	4
Neighborhood	Old East York
Crime_Count	479
Distance_From_Office	2.99
Category	45
Cluster Labels	2
1st Most Common Venue	Coffee Shop
2nd Most Common Venue	Park
3rd Most Common Venue	Indian Restaurant
4th Most Common Venue	Thai Restaurant
5th Most Common Venue	Bus Stop
6th Most Common Venue	American Restaurant
7th Most Common Venue	Pub
8th Most Common Venue	Breakfast Spot
9th Most Common Venue	Italian Restaurant
10th Most Common Venue	Metro Station
11th Most Common Venue	Shopping Mall
12th Most Common Venue	Playground NA
13th Most Common Venue	Sports Bar NA
14th Most Common Venue	Cocktail Bar NA
15th Most Common Venue	Shopping Plaza NA
16th Most Common Venue	Buffet NA
17th Most Common Venue	Big Box Store NA

CONCLUSION

Considering the scope of the criteria and the available data, I think the implemented model performed quite well. As people like my brother moves a lot now-a-days, it would have been very helpful to have this kind of models in place.

For this study, I have only included Crime data and venues (facilities). We can make this model more efficient, if we can consider other below data sources as well:

- Education specific data sources
- Employment specific data sources
- Medical Health support facilities specific data sources
- Available residential communities with price index data source