Neighborhood Recommendation

City of Toronto

Business Problem

- 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.

- 140 neighborhoods, which one to choose

Roadmap

Business requirement
- Define requirement
- Capture different criteria

Data Acquisition
- Identify proper data sources
- Collect data sources
- Apply data cleansing
- Apply necessary transformation

Methodology
- Select appropriate features
- Exploratory Data Analysis
- Data Visualization
- Model Creation

Result
- Explain the findings
- If further data required then go back to Step 2

Discussion and Conclusion
- Recommend neighborhood
- Explain reasons for decision
- Conclude the model

Iterative Process

Business requirements

Facility Preferences

- Safe Neighborhood

- Coffee Shops

- Restaurants (Indian, Italian, Thai and American)

- Transportation - Metro Station, Bus Station

- Breakfast places

- Selected Bars - Sports Bar, Cocktail Bar, Pub

- Outdoor Activity Center - Playground, Park

- Shopping Center

Additional Criteria

- Not more than 6 miles away from Office

- Shopping Center is not a for daily need, so it is good to have but not mandatory

Data Acquisition – Identification of data sources

Toronto Neighborhood boundary data

- Source: Toronto Department of Police
- Link

Toronto Major Criminal Incident data

- Source: Toronto Department of Police
- Link

Venues near Toronto neighborhoods
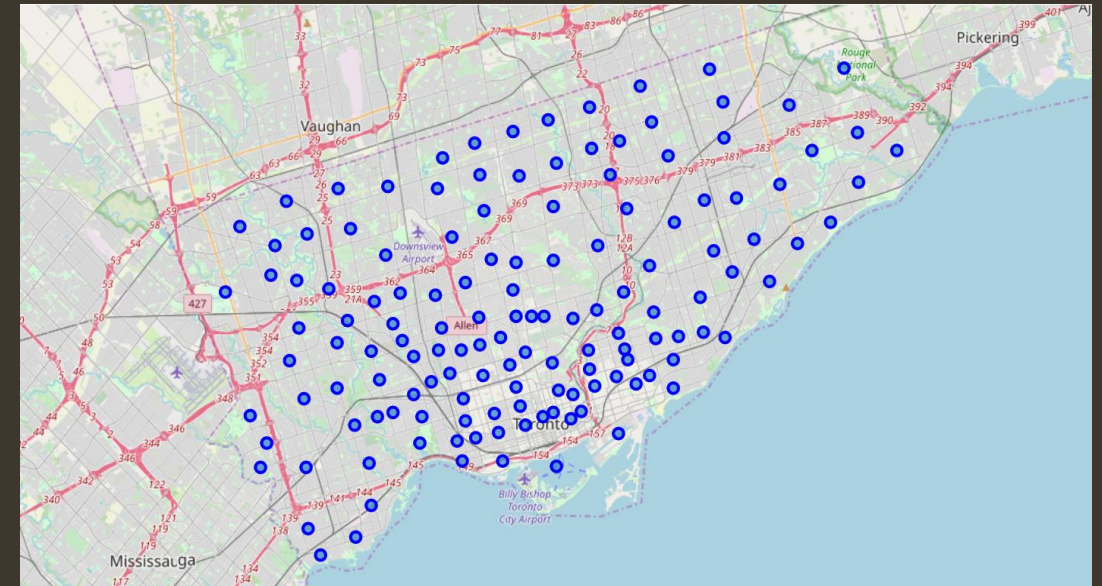
- Source: Foursquare API

Methodology

Selection of features from Toronto Neighborhood boundary data

- This data source provides below information for each of the 140 neighborhoods:
  - Average crime data starting from 2014 till 2018
  - Name of the Neighborhood
  - Boundary coordinates

- We will extract the below information from this dataset:
  - Drop Average crime data starting from 2014 till 2018, as it is old
  - Use Name of the Neighborhood
  - Use Boundary coordinates
  - Calculating the centroid (Latitude and Longitude) of each neighborhood using boundary coordinates
  - Calculate the distance of Office from the centroid of the neighborhood

Methodology

Selection of features from Toronto Neighborhood boundary data

Methodology

Selection of features from Toronto Major Criminal Incidents data

- This data source provides below information for each of the 140 neighborhoods:
  - All major criminal incidents starting from 2014 till 2019 with type of crime
  - Name of the Neighborhood
  - Coordinates where the incident happened

- We will extract the below information from this dataset:
  - Drop Type of crime and other crime specific information
  - Use Name of the Neighborhood
  - Calculate number of criminal activities per neighborhood
  - Percentage of Criminal activities per neighborhood

Methodology

Combined Data Source: Merge crime data with neighborhood geographic information

- Name of neighborhood

- Coordinate of centroid (Latitude, Longitude)

- Total number of criminal incidents

- Percentage of Crime

- Distance from Office

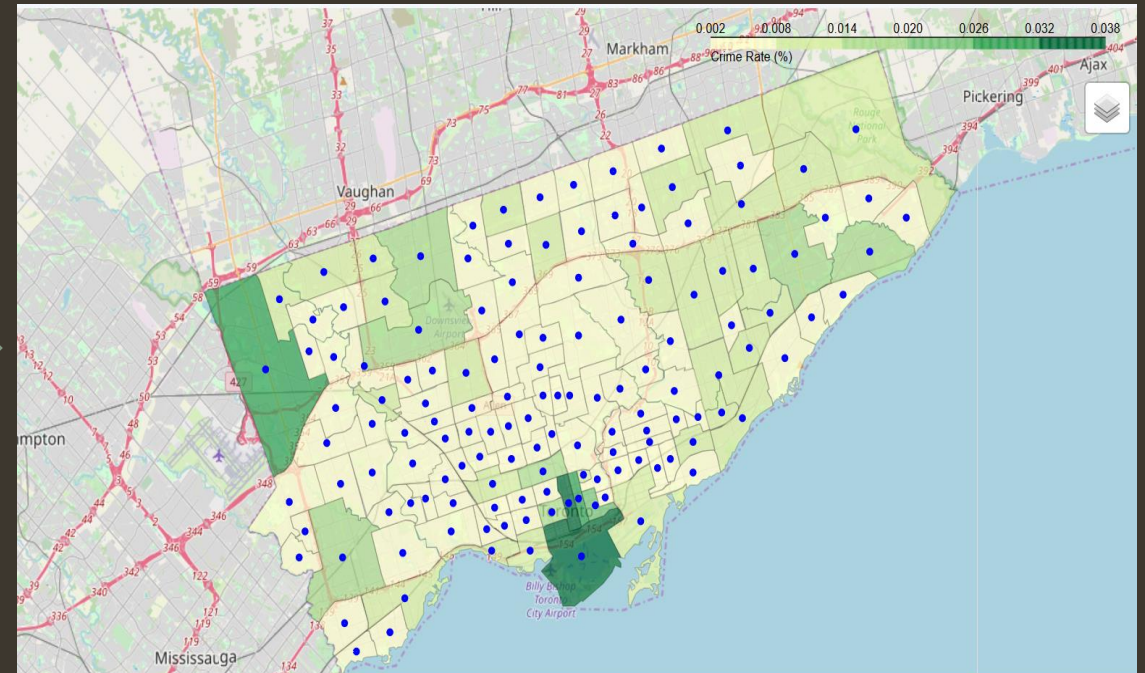| Combined Data sorted based on total criminal incidents					
	Neighborhood	Crime_Count	Percentage_Crime	Latitude	Longitude	Distance_From_Office
0	Lambton Baby Point	353	0.001710	43.657421	-79.496008	5.72
1	Woodbine-Lumsden	377	0.001826	43.694107	-79.311123	3.96
2	Maple Leaf	410	0.001986	43.715575	-79.480718	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

Methodology

Combined Data Source: Merge crime data with neighborhood geographic information

	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

Combined Data sorted based on total criminal incidents

Methodology

Filter first 100 neighborhoods after sorting based on criminal incident

First 100 neighborhood dataset based on total criminal incidents

	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

```
safe_neighborhood_with_crime_df.shape

(100, 6)
```
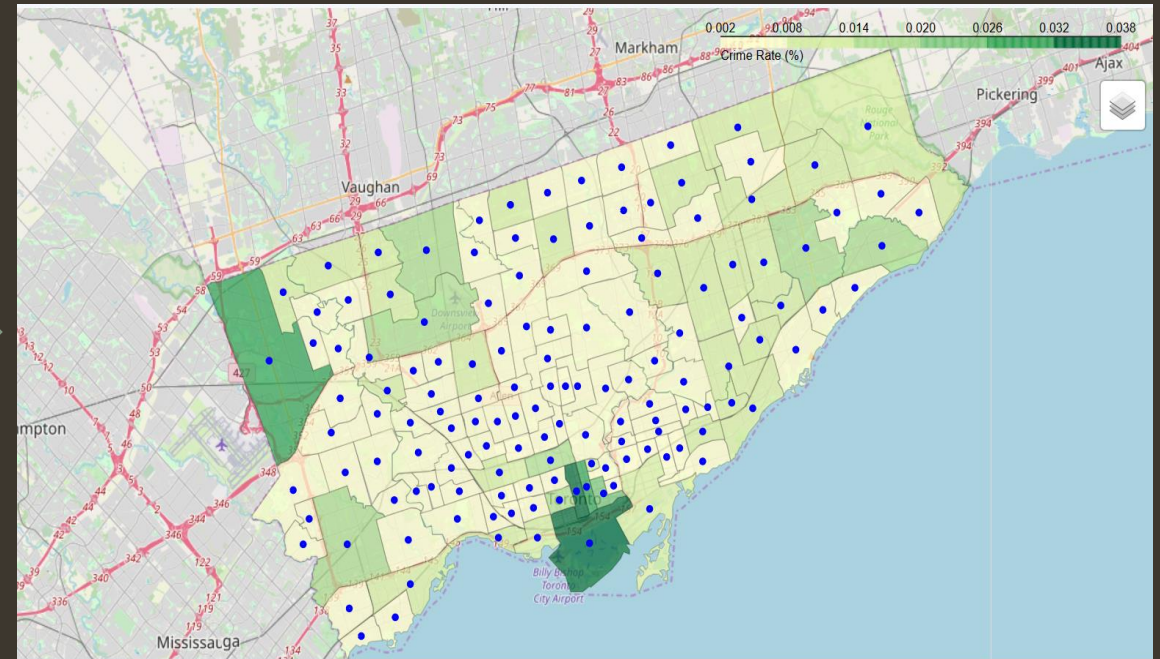
Methodology

Get all the neighborhoods which are within 6 miles away from office

Neighborhood dataset based on distance filter

	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	Yonge-St.Clair	412	0.001996	43.687859	-79.397831	1.36
4	Old East York	479	0.002320	43.696781	-79.335448	2.99

```
distance_based_safe_neighborhood_df.shape
```

```
(59, 6)
```
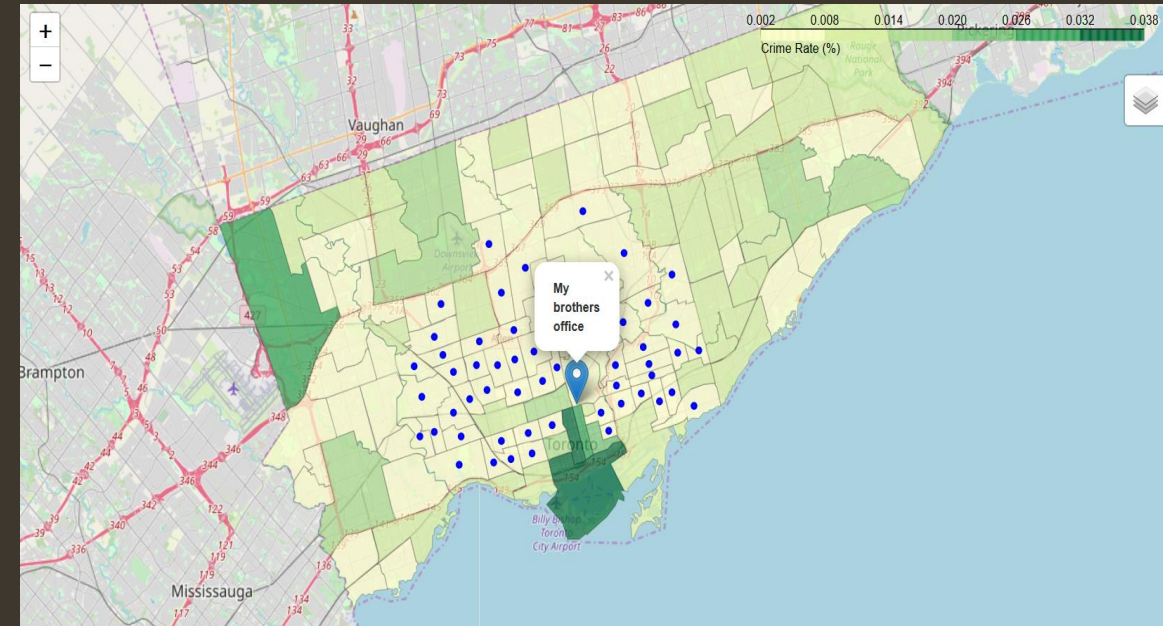
We will proceed with these 59 neighborhood for further analysis with the neighborhood data. As safety is a major criteria, later on, we will apply another data filter based on the coordinates of the venues.

Methodology

Data Source: Foursquare API

- Collect all the category ids for each preferred categories my brother has mentioned. Use Foursquare Category hierarchy page for this.

- Use Foursquare API for search venues to retrieve all the preferred venues which are within 1500 meters radius of the centroid of each of 59 neighborhoods.

- Prepare a dataset with the below information:

  - Name of the Venue

  - Category of the Venue

  - Coordinates of the Venue

Methodology

Retrieve all preferred venues within 1500 meters of neighborhood centroid

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

Methodology

Identification of unsafe venues and removing them

- Collect the boundary information of all the unsafe neighborhoods (40)

- For each venue out of retrieved 2542 venues:
  - Check if the coordinate of the venue resides within the boundary of any unsafe neighborhood
  - If not, then consider the venue, otherwise, ignore the venue

- Python's shapely package is used for this activity

```python
def checkIfCoordinateInNeighborhoodBoundary(point_latitude, point_longitude, neighborhood_geometry):
    polygon = Polygon(neighborhood_geometry)
    target_point = Point(point_longitude, point_latitude)
    return target_point.within(polygon)
```

Methodology

Filter out all the venues which are geographically located with unsafe neighborhoods

Venue dataset based on safe neighborhood

	Neighborhood	Venue_Name	Category	Venue_Latitude	Venue_Longitude
0	Lambton Baby Point	Magwood Park	Park	43.657954	-79.498094
1	Lambton Baby Point	Jane Subway Station	Metro Station	43.649944	-79.483995
2	Lambton Baby Point	Walmart	Big Box Store	43.668332	-79.485117
3	Lambton Baby Point	Park Lawn Park	Park	43.642014	-79.491153
4	Lambton Baby Point	Tim Hortons	Coffee Shop	43.667785	-79.488055
5	Lambton Baby Point	Strada	Italian Restaurant	43.650565	-79.478567
6	Lambton Baby Point	La Veranda Osteria	Italian Restaurant	43.647638	-79.511442
7	Lambton Baby Point	Royal York Subway Station	Metro Station	43.648119	-79.511381
8	Lambton Baby Point	Old Mill Subway Station	Metro Station	43.649892	-79.495322
9	Lambton Baby Point	Starbucks	Coffee Shop	43.648350	-79.507680

```
filtered_neighborhood_venue_df.shape

(2032, 5)
```
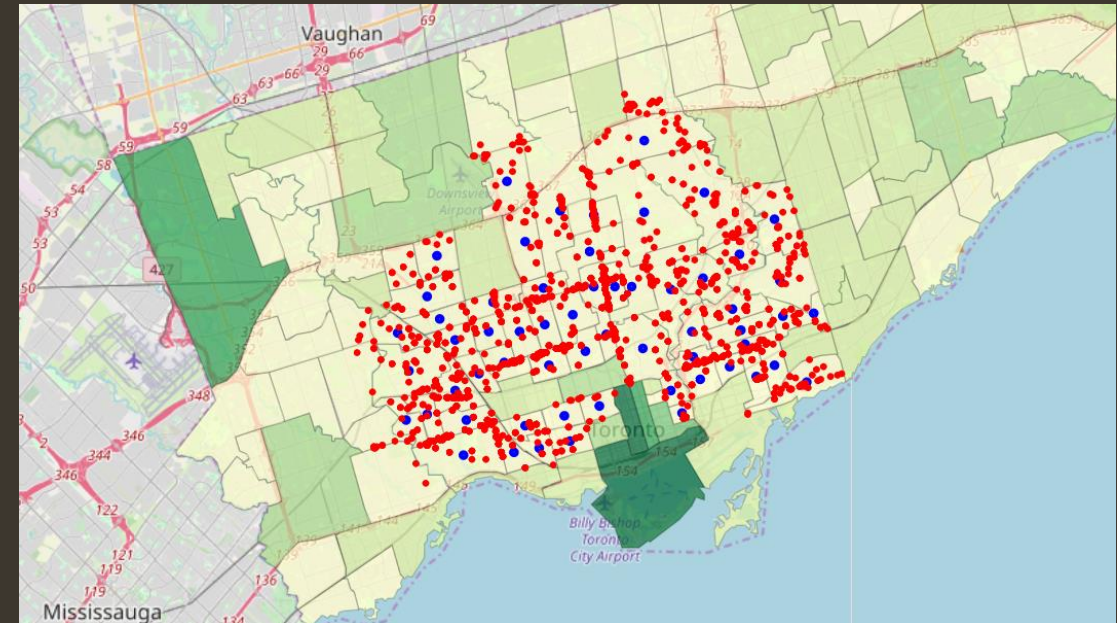

Now the number of venues got dropped from 2542 to 2032.

Methodology

Preparation for K-mean clustering

- We have total 17 unique venue categories

- Apply one hot encoding to transform categorical feature into numerical feature

- Transform the dataset by applying group by function on each neighborhood and calculating mean of each venue category

	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	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 we are ready to feed this data to K-mean clustering function. After some repetitive process, I have decided to create 3 clusters.

Results

Plot the cluster labeled neighborhoods

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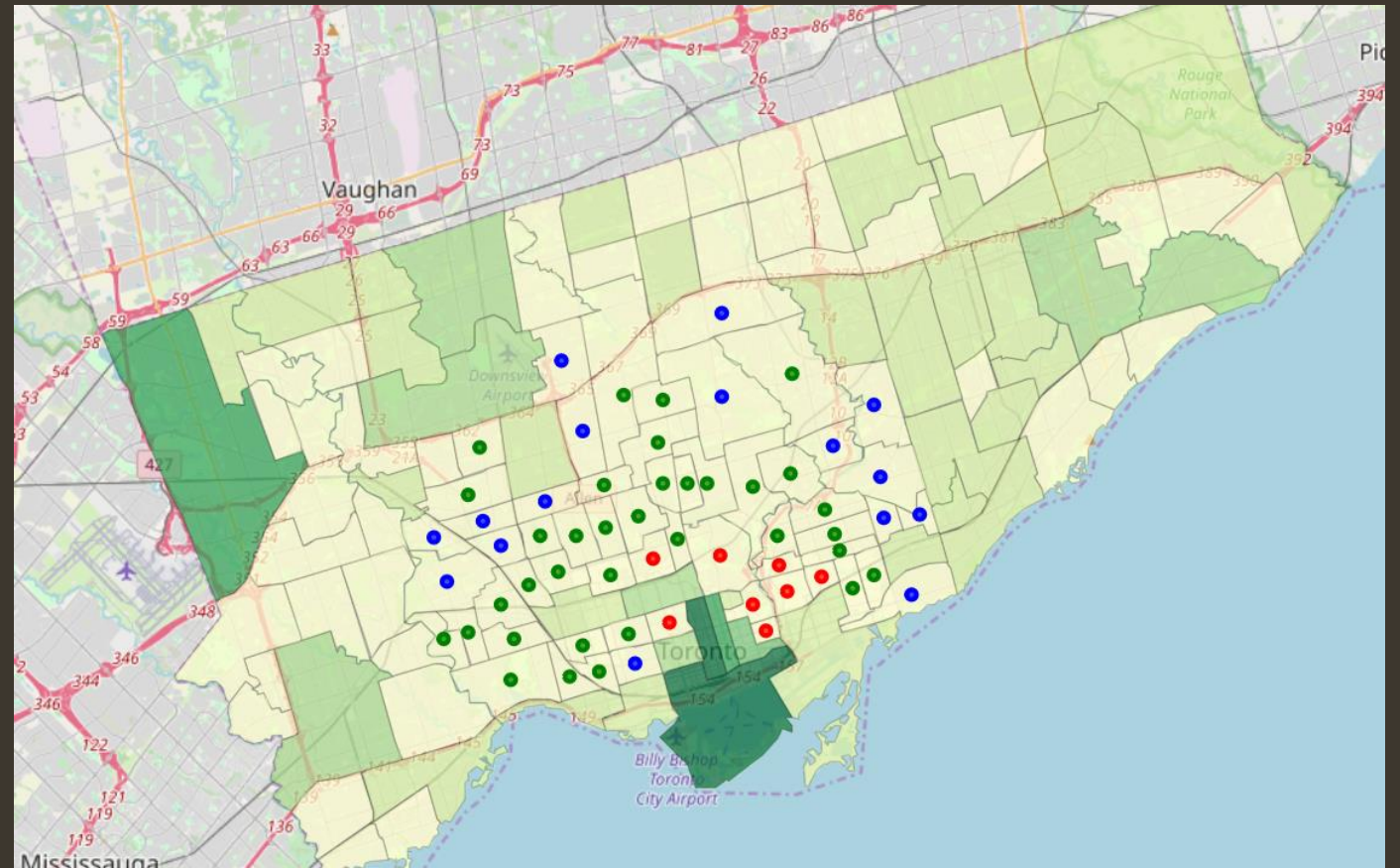
Discussions

- 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.
- Cluster 1 (BLUE in color). These neighborhoods are also adjacent to other unsafe neighborhoods which forces the preferred venues to ignore.
- Cluster 2 neighborhoods (GREEN in color) are the safest neighborhoods and also surrounded by safe neighborhoods.

Recommended neighborhoods from Cluster 2

1. "Lambton Baby Point", because it is having the lowest criminal incidents reported. It also has almost all venues my brother had mentioned.
2. "Old East York" which has "Indian Restaurant as the third popular venue and also it is closer to his office.

Discussions

Popular venues at the 2 recommended neighborhoods

	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

Conclusions

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

Add a Slide Title - 4