Capstone Project - The Battle of the Neighborhoods (Week 2)

Applied Data Science Capstone by IBM/Coursera

Table of contents

Introduction: Business Problem

In this project we will try to find an optimal location to live for an immigrant, who doesn't know anything about whereabouts of the area. Specifically, this report will be **targeted to immigrants service officers or NGO's, who helps immigrant to settle down**. Lots of people are migrating to various states of Canada and needed lots of research for good neiborhood as well as better facilities around. This project is for those people who are looking for better neighborhoods to ease of accessing to Cafe, School, Super market, medical shops, grocery shops, mall, theatre, hospital, like minded people, etc.

Since there are lots of look for and a person would like to settle a place, where they can find good connection with daily needs. We are definately looking for the area, that is safe and crime free and also has marketplace nearby.

The purpose of this Project is to help people in exploring better facilities around their neighborhood. It will help people making smart and efficient decision on selecting great neighborhood out of numbers of other neighborhoods in **North York, Toranto.** This Project also aims to create an analysis of features for a people migrating to Canada to search a best neighborhood as a comparative analysis between neighborhoods. The features include **crime rates in neighborhood** in neighborhood.

It will help people to get awareness of the area and neighborhood before moving to a new city, state, country or place for their work or to start a new fresh life. We will use our data science powers to generate a few most promissing neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen.

Data

Data Link 1: [https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
Data Link 2: [https://open.toronto.ca/dataset/neighbourhood-crime-rates/ (https://open.toronto.ca/dataset/neighbourhood-crime-rates/)](https://open.toronto.ca/dataset/neighbourhood-crime-rates/)

I'll use North York dataset, which we have already scrapped from wikipedia on Week 3. Dataset consisting of postal codes, borough, neighborhoods, latitude and longitude in csv file name 'Canada.csv'

Other than that I'll also look for Canada neighborhood crime rates in data to find the better neighborhood for immigrants. For each neighborhood, we have chosen **the radius to be 100 meter.

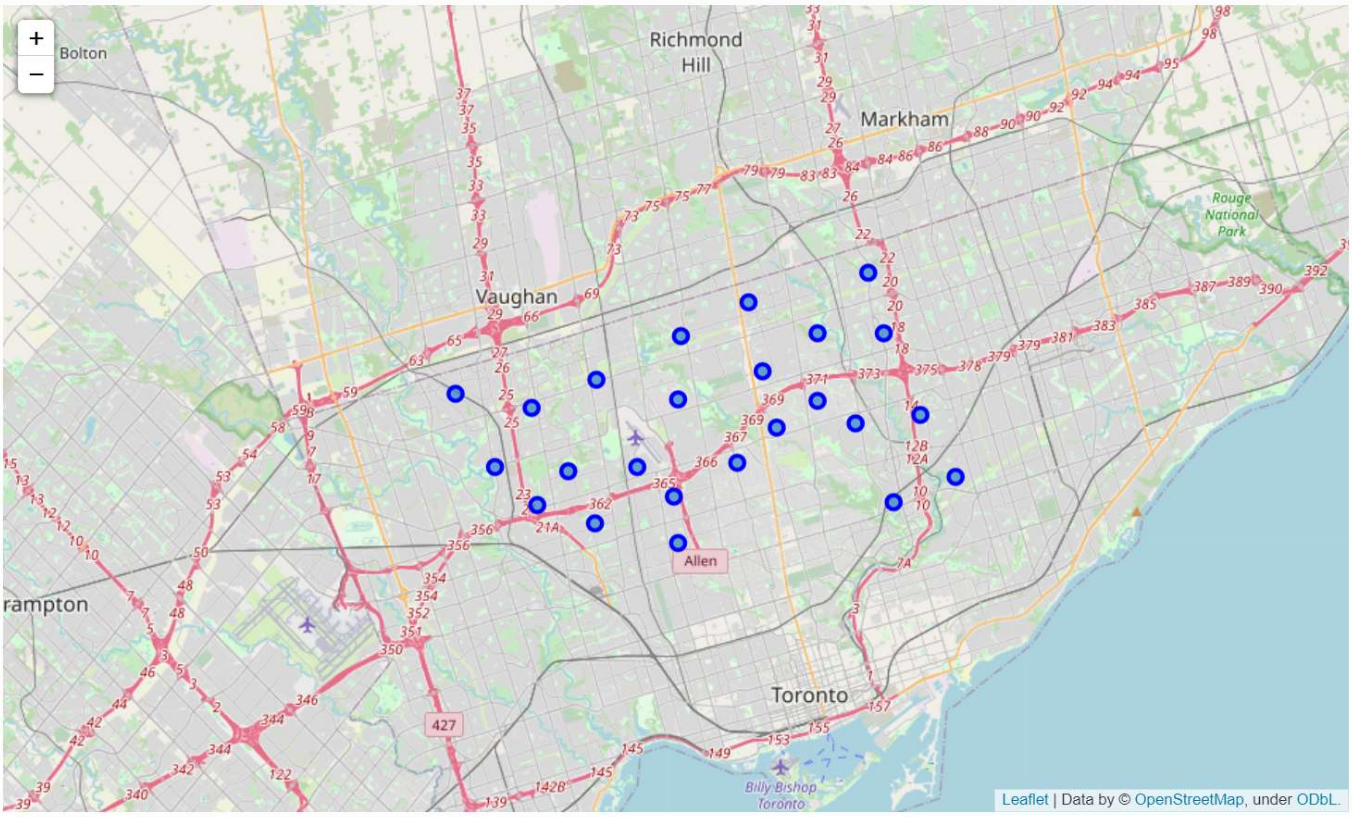
Based on definition of our problem, factors that will influence our decission are:

- neighborhood in North York
- Crime rate in that area
- Neighborhood
- Neighborhood Latitude
- Neighborhood Longitude
- Venue
- Name of the venue e.g. the name of a store or restaurant
- Venue Latitude
- Venue Longitude
- Venue Category

Following data sources will be needed to extract/generate the required information:

- Foursquare API location provider technology to obtain **Google Maps API reverse geocoding**
- Points of interest near by the location in every neighborhood will be obtained using **Foursquare API**
- Segmentation and clustering of data with **K-Means Clustering algorithm**

North York, Canada



Neighborhood Candidates

Let's create latitude & longitude coordinates for centroids of our candidate neighborhoods. We will create a grid of cells covering our area of interest which is aprox. 100 meter around North york.

Let's first find the latitude & longitude

In [53]:

```
# Adding Columns Latitude & Longitude
df_coord = pd.DataFrame(coord, columns=['Latitude', 'Longitude'])
df['Latitude'] = df_coord['Latitude']
df['Longitude'] = df_coord['Longitude']

df[df.PostalCode == 'M5G']
```

Out[53]:

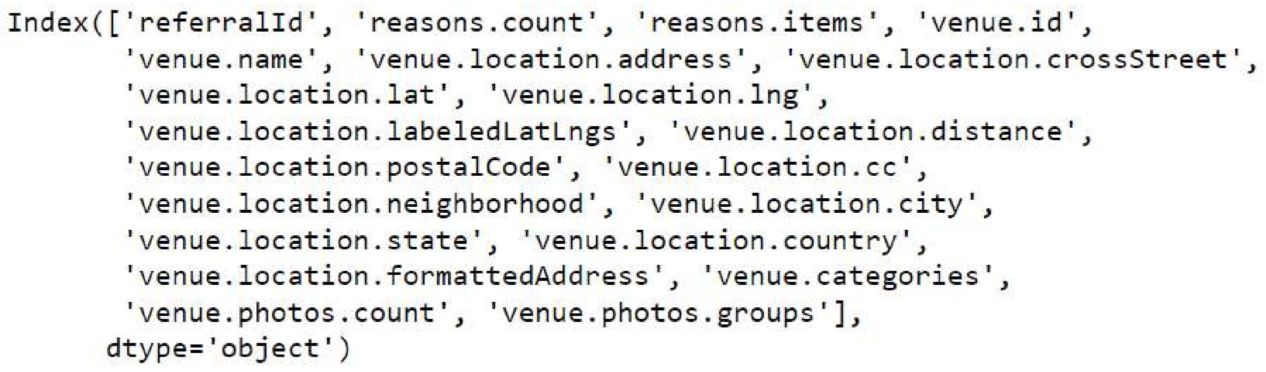
	PostalCode	Borough	Neighborhood	Latitude	Longitude
24	M5G	Downtown Toronto	Central Bay Street	43.65609	-79.38493

Foursquare

Foursquare is a location data provider with information of venues and events with info like ; location and even photos in an area of interest. We will use data for different venues in different neighborhoods of that specific borough.For achieving that information, we'll use "Foursquare" locational provider. Now that we have our location candidates, let's use Foursquare API to get info on poin of interest in each neighborhood.

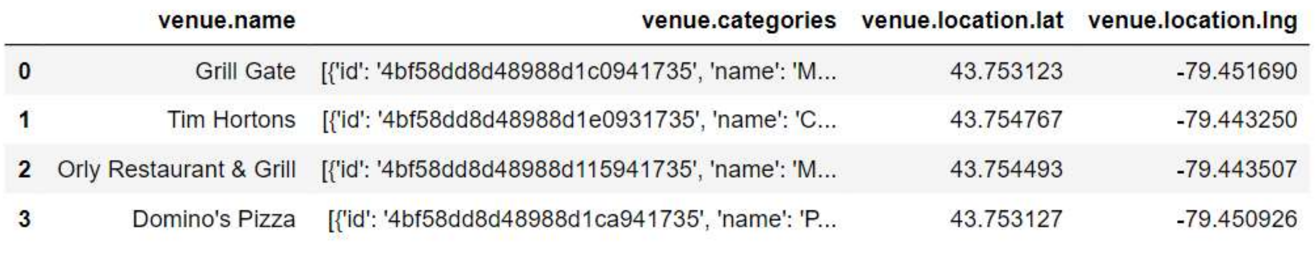
Foursquare credentials are defined in hidden cell bellow.

radius = 500 LIMIT = 100 url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(CLIENT_ID, CLIENT_SECRET, VERSION, latitude, longitude, radius, LIMIT) results = requests.get(url).json()venues = results['response']['groups'][0]['items'] nearby = json_normalize(venues) nearby.columns



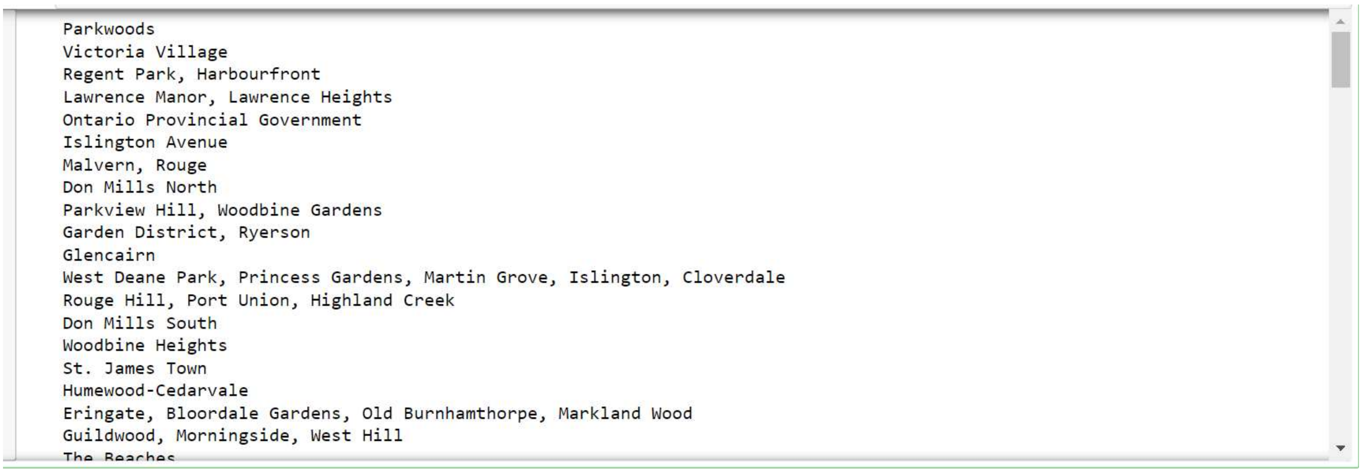
```
Index(['referralId', 'reasons.count', 'reasons.items', 'venue.id',
       'venue.name', 'venue.location.address', 'venue.location.crossStreet',
       'venue.location.lat', 'venue.location.lng',
       'venue.location.labeledLatLngs', 'venue.location.distance',
       'venue.location.postalCode', 'venue.location.cc',
       'venue.location.neighborhood', 'venue.location.city',
       'venue.location.state', 'venue.location.country',
       'venue.location.formattedAddress', 'venue.categories',
       'venue.photos.count', 'venue.photos.groups'],
      dtype='object')
```

filtered_column = ['venue.name', 'venue.categories', 'venue.location.lat', 'venue.location.lng'] nearby =nearby.loc[:, filtered_column] nearby.head()



	venue.name	venue.categories	venue.location.lat	venue.location.lng
0	Grill Gate	[{'id': '4bf58dd8d48988d1c0941735', 'name': 'M...	43.753123	-79.451690
1	Tim Hortons	[{'id': '4bf58dd8d48988d1e0931735', 'name': 'C...	43.754767	-79.443250
2	Orly Restaurant & Grill	[{'id': '4bf58dd8d48988d115941735', 'name': 'M...	43.754493	-79.443507
3	Domino's Pizza	[{'id': '4bf58dd8d48988d1ca941735', 'name': 'P...	43.753127	-79.450926

NorthYork_venues = get_Venues(names=df_2['Neighborhood'], latitudes=df_2['Latitude'], longitudes=df_2['Longitude'])



```
Parkwoods
Victoria Village
Regent Park, Harbourfront
Lawrence Manor, Lawrence Heights
Ontario Provincial Government
Islington Avenue
Malvern, Rouge
Don Mills North
Parkview Hill, Woodbine Gardens
Garden District, Ryerson
Glencairn
West Deane Park, Princess Gardens, Martin Grove, Islington, Cloverdale
Rouge Hill, Port Union, Highland Creek
Don Mills South
Woodbine Heights
St. James Town
Humewood-Cedarvale
Eringate, Bloordale Gardens, Old Burnhamthorpe, Markland Wood
Guildwood, Morningside, West Hill
The Beaches
```

Looking good. So now we have a lot of details in area within 100 meters area. This concludes the data gathering phase - we're now ready to use this data for analysis to produce the report on optimal locations to choose.

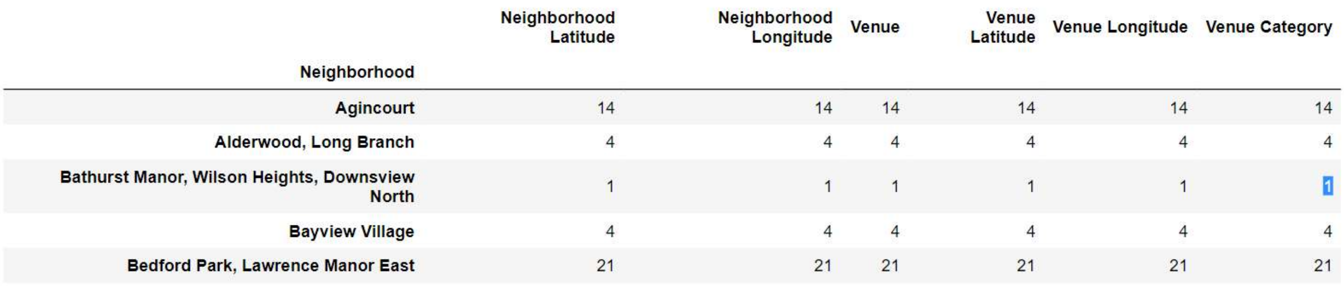
Methodology

Clustering Approach: In this project we will direct our efforts on detecting areas of North York. To compare the similarities between two cities, we decided to explore neighborhoods, segmented them, and grouped them into clusters to find similar neighborhoods in a big city like New York and Toronto. To be able to do that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm. We limited our analysis to area 100 meter around city center.

In first step we have collected the required **data: location, langitude, latitude, number of crimes in the area of every neighborhood**. We have also **identified point of interest nearby location** (according to Foursquare categorization).

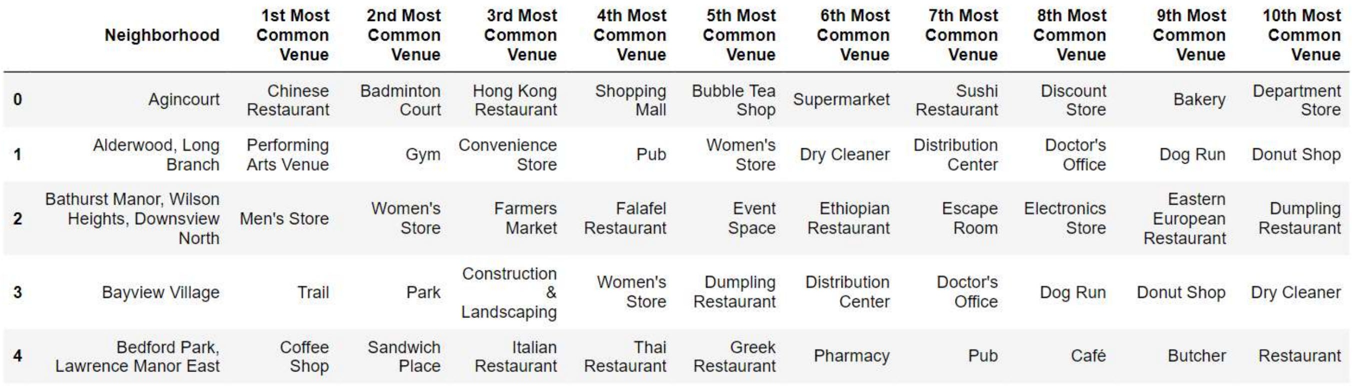
Second step in our analysis will be calculation and exploration of most common point of interest across different areas of North York.

```
There are 261 Uniques Categories.
```



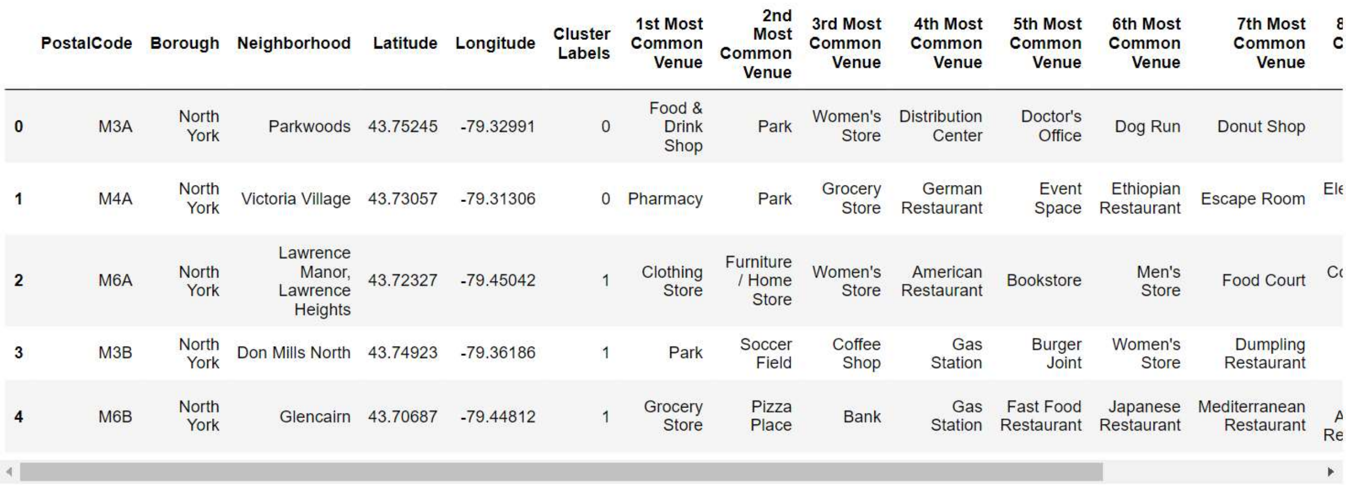
	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Agincourt	14	14	14	14	14	14
Alderwood, Long Branch	4	4	4	4	4	4
Bathurst Manor, Wilson Heights, Downsview North	1	1	1	1	1	1
Bayview Village	4	4	4	4	4	4
Bedford Park, Lawrence Manor East	21	21	21	21	21	21

In third and final step we will focus on most promising areas and within those create **clusters of locations that meet some basic requirements**: we will take into consideration locations with **markets, schools or other facilities around in radius of 100 meters**, and we also want locations that has **crime free neighborhood**. We will present map of all such locations but also create clusters (using **k-means clustering**) of those locations to identify neighborhoods / langitude / longitude which should be a starting point for final 'street level' exploration and search for optimal venue location.



	Neighborhood	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	Agincourt	Chinese Restaurant	Badminton Court	Hong Kong Restaurant	Shopping Mall	Bubble Tea Shop	Supermarket	Sushi Restaurant	Discount Store	Bakery	Department Store
1	Alderwood, Long Branch	Performing Arts Venue	Gym	Convenience Store	Pub	Women's Store	Dry Cleaner	Distribution Center	Doctor's Office	Dog Run	Donut Shop
2	Bathurst Manor, Wilson Heights, Downsview North	Men's Store	Women's Store	Farmers Market	Falafel Restaurant	Event Space	Ethiopian Restaurant	Escape Room	Electronics Store	Eastern European Restaurant	Dumpling Restaurant
3	Bayview Village	Trail	Park	Construction & Landscaping	Women's Store	Dumpling Restaurant	Distribution Center	Doctor's Office	Dog Run	Donut Shop	Dry Cleaner
4	Bedford Park, Lawrence Manor East	Coffee Shop	Sandwich Place	Italian Restaurant	Thai Restaurant	Greek Restaurant	Pharmacy	Pub	Café	Butcher	Restaurant

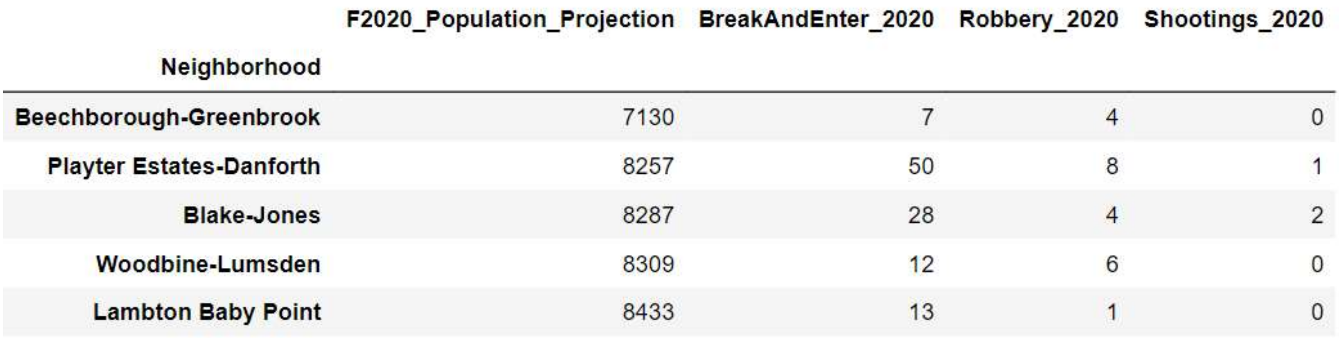
Clustered data by K- Means



	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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	8 C
0	M3A	North York	Parkwoods	43.75245	-79.32991	0	Food & Drink Shop	Park	Women's Store	Distribution Center	Doctor's Office	Dog Run	Donut Shop	
1	M4A	North York	Victoria Village	43.73057	-79.31306	0	Pharmacy	Park	Grocery Store	German Restaurant	Event Space	Ethiopian Restaurant	Escape Room	Ele
2	M6A	North York	Lawrence Manor, Lawrence Heights	43.72327	-79.45042	1	Clothing Store	Furniture / Home Store	Women's Store	American Restaurant	Bookstore	Men's Store	Food Court	Co
3	M3B	North York	Don Mills North	43.74923	-79.36186	1	Park	Soccer Field	Coffee Shop	Gas Station	Burger Joint	Women's Store	Dumpling Restaurant	
4	M6B	North York	Glencairn	43.70687	-79.44812	1	Grocery Store	Pizza Place	Bank	Gas Station	Fast Food Restaurant	Japanese Restaurant	Mediterranean Restaurant	A Re

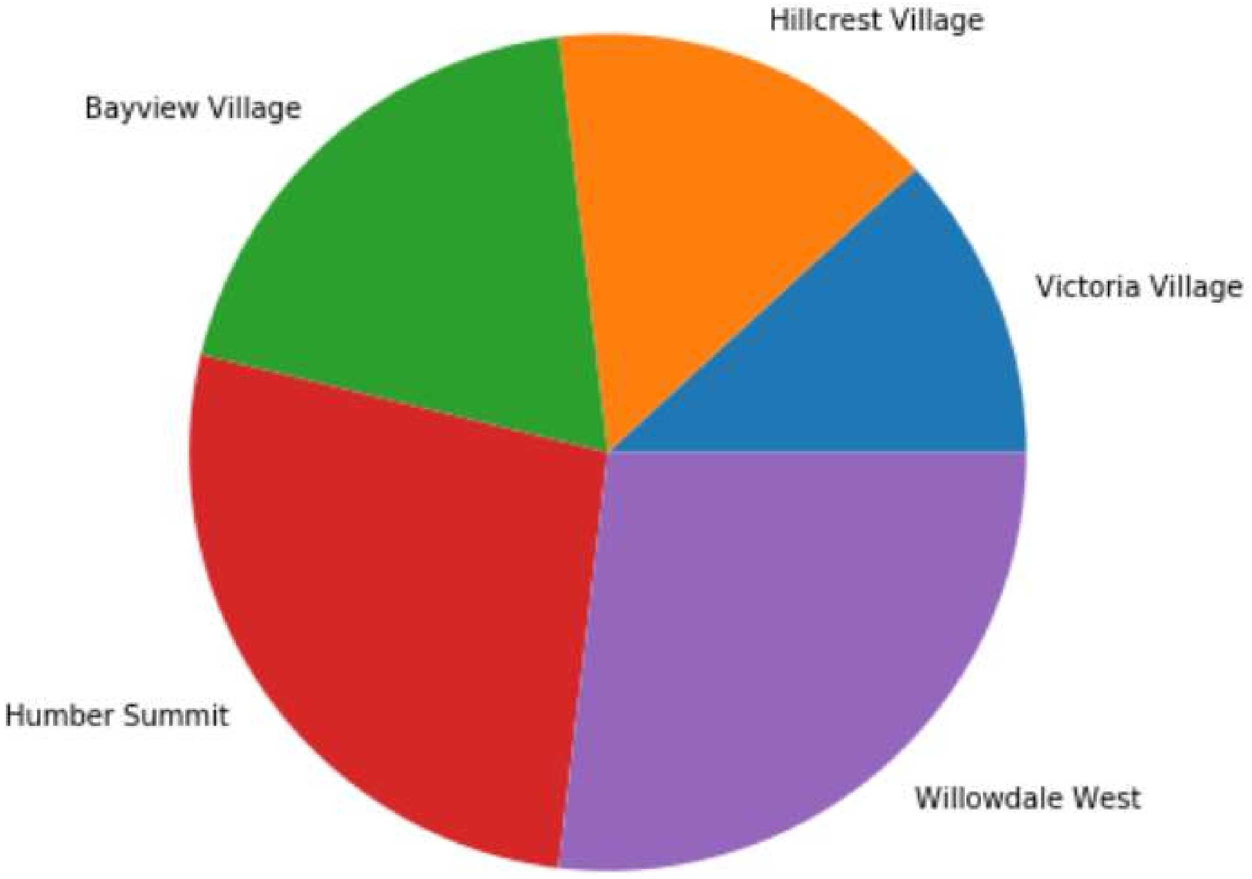
Analysis

Let's perform some basic explanatory data analysis and derive some additional info from our raw data. First let's count the **crime neighborhood in North York**:



	F2020_Population_Projection	BreakAndEnter_2020	Robbery_2020	Shootings_2020
Neighborhood				
Beechborough-Greenbrook	7130	7	4	0
Playter Estates-Danforth	8257	50	8	1
Blake-Jones	8287	28	4	2
Woodbine-Lumsden	8309	12	6	0
Lambton Baby Point	8433	13	1	0

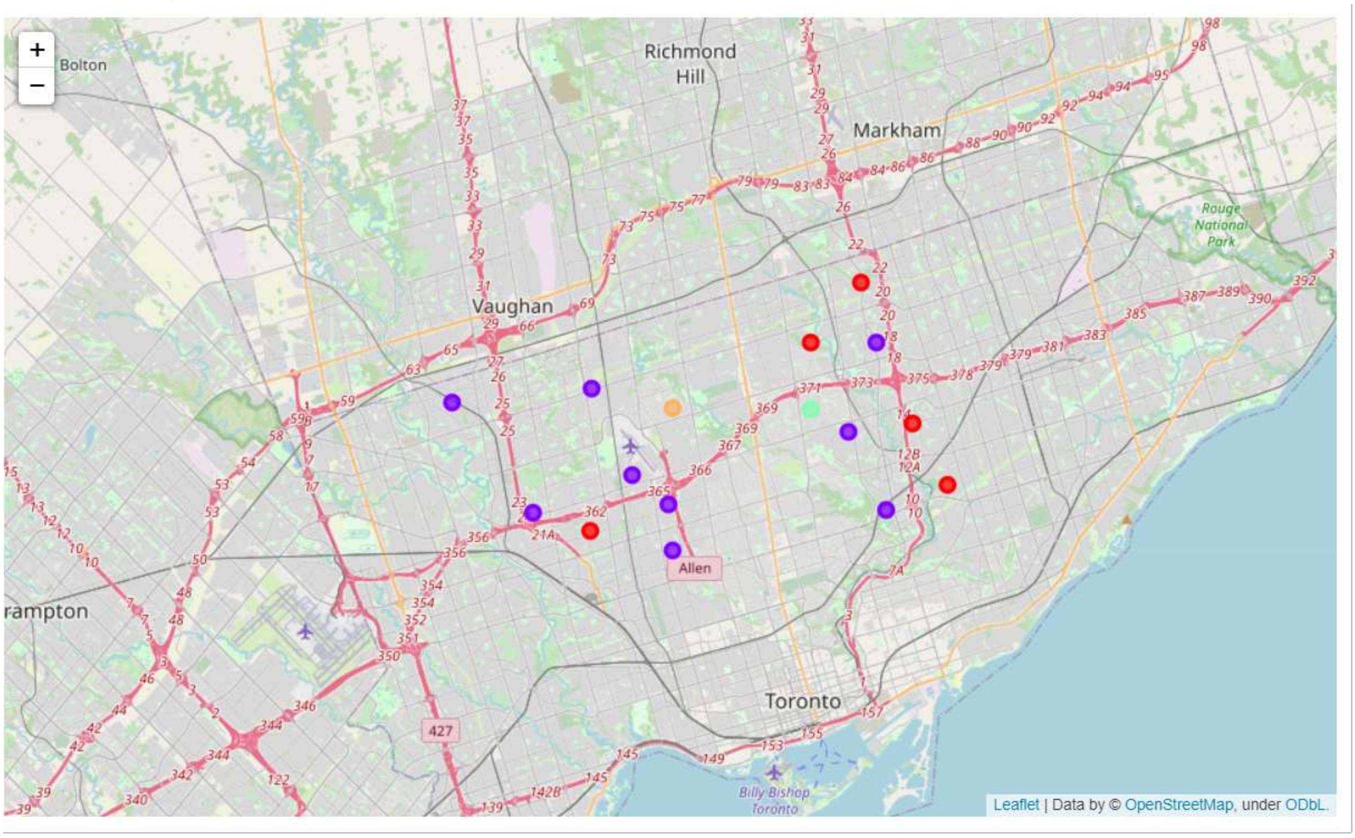
**'Robbery_2020'

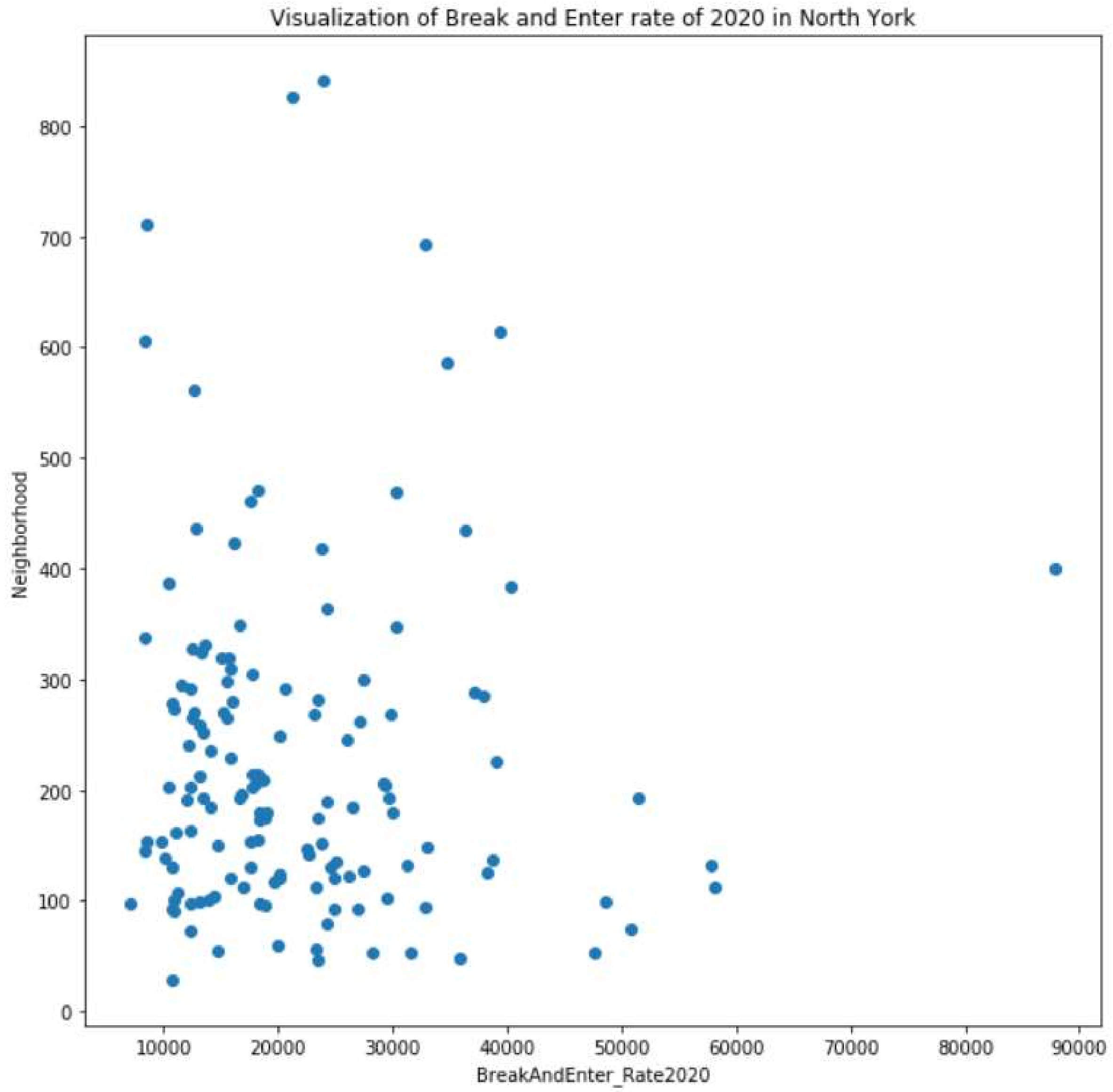


OK, now let's calculate the **distance to nearest Italian restaurant from every area candidate center** (not only those within 300m - we want distance to closest one, regardless of how distant it is).

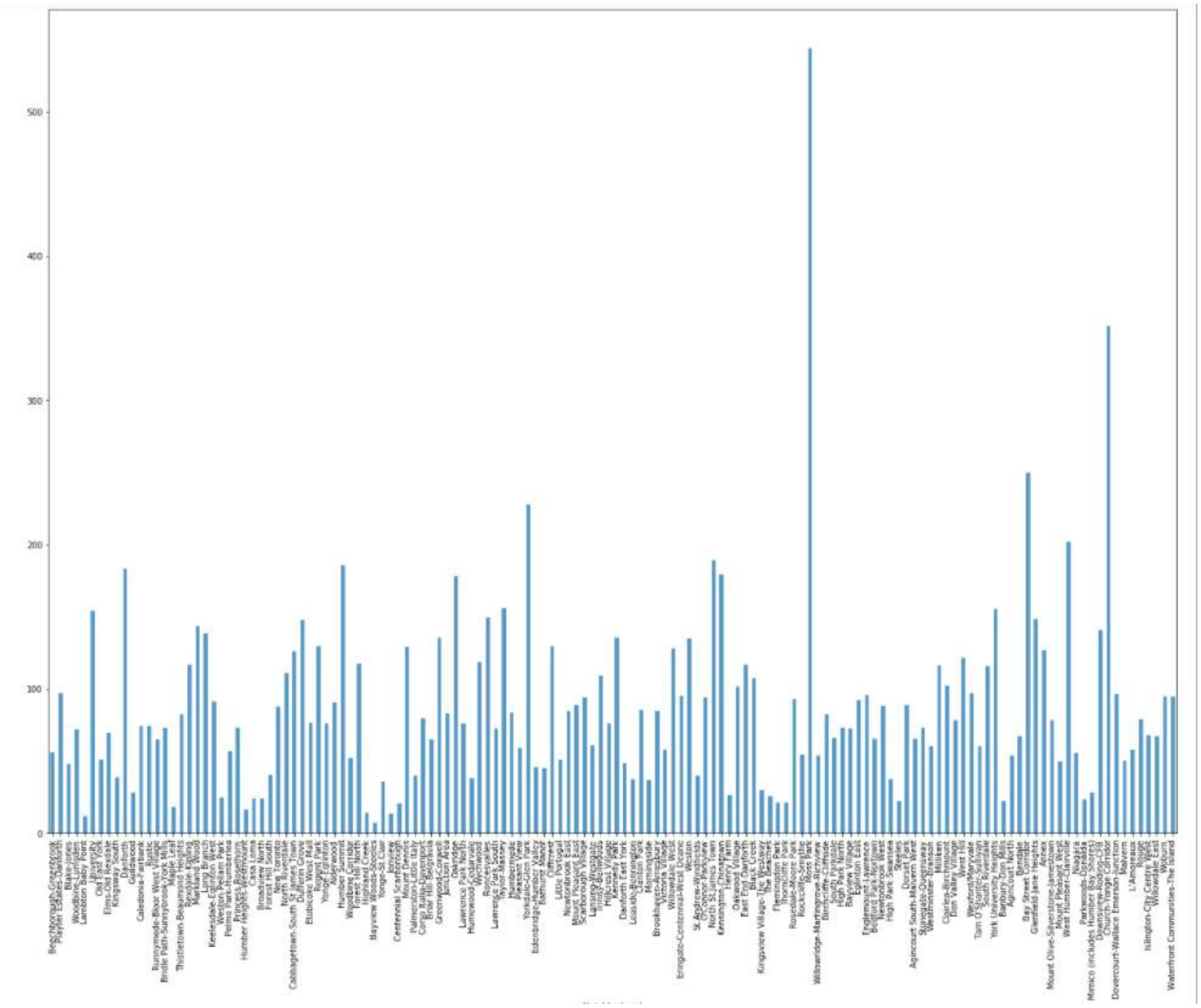
Results and Discussion

Clustered map of North York




Visualization of Break and Enter rate of 2020 in North York
Neighborhood
0
100
200
300
400
500
600
700
800
10000
20000
30000
40000
50000
60000
70000
80000
90000
BreakAndEnter_Rate2020

Robbery rate of North York in 2020



Our analysis shows that although there is a big neighbourhood in North York, the more area of interest is in Cluster -2 (Lawrence Manor, Lawrence Heights, Don Mills North, Glencairn, Don Mills South, Fairview, Henry Farm, Oriole, Northwood Park, York University, Downsview East and Downsview West), and these areas are also crime free and very close to downtown, which offer a combination of popularity among tourists, closeness to city center, strong socio-economic dynamics.

K-Means Clustered: Those location candidates were then clustered to create zones of interest which contain greatest number of location candidates. Addresses of centers of those zones were also generated using reverse geocoding to be used as markers/starting points for more detailed local analysis based on other factors.

Foursquare API: This project have used Four-square API as its prime data gathering source as it has a database of millions of places, especially their places API which provides the ability to perform location search, location sharing and details about a business.

Conclusion

Purpose of this project was to identify best suited areas close to center with low number of crime in order to find a positive starting in a country. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration.

In this project, using k-means cluster algorithm I separated the neighborhood into 10(Ten) different clusters and for more than 100 different lattitude and logitude from dataset, which have very-similar neighborhoods around them. Using the charts above results presented to a particular neighborhood based on crime rate in the neighbourhood. 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 financial 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.

Libraries Which are Used to Develope the Project:

1. Pandas: For creating and manipulating dataframes.
2. Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.
3. Scikit Learn: For importing k-means clustering.
4. JSON: Library to handle JSON files.
5. XML: To separate data from presentation and XML stores data in plain text format.
6. Geocoder: To retrieve Location Data.
7. Beautiful Soup and Requests: To scrap and library to handle http requests.
8. Matplotlib: Python Plotting Module.

In [ ]:

```

```