Peer-graded Assignment: Capstone Project - The Battle of Neighborhoods

Introduction/Business Problem

I have been approached by a group of professional couples (Pharmacists and Engineers) who intent to emigrate to the United States or Canada from Nigeria. They intend to live in Brooklyn, New York or Toronto, Canada.

The Electrical Engineer believes he will be comfortable finding a job in any of the locations but the pharmacists requested that I compare between the two locations and their neighborhood in order to find a suitable location for the pharmacy outlet while maintaining a relatively calm family life.

Our terms of reference are as follows:

1. To compare between Toronto and Brooklyn New York City in order to decide on what city is preferred for an emigrant from Nigeria in terms of cultural diversity and tolerance, proximity to relaxation and recreational centers, and relatively serene neighborhoods.
2. After choosing between Toronto and New York, I will then decide on best neighborhood to situate a pharmacy outlet using data analysis. My choice of neighborhoods to situate a pharmacy outlet will be based on neighborhoods with already high "most common venues visited" as pharmacy as that will mean that there is room for competition in the market share in such neighborhoods

Data section

To achieve the above business problem, I will be using data from the following links:

1. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
   This Wikipedia site contains an updated neighborhoods and boroughs in the Toronto metropolis. It is in an unstructured data format and hence I used BeautifulSoup to extract it

2. http://cocl.us/Geospatial_data
   This is an already extracted gps coordinates of neighborhood in Toronto
3. Foursquare (www.foursquare.com): This is a technology company that has a large dataset of accurate location data. It is widely used by top companies such as Apple, Uber, Twitter, and Snapchat. Their API is currently being used by over 100,000 developers
4. newyork_data.json from https://cocl.us/new_york_dataset: This data set gives us the new York neighborhood inclusive of their gps coordinates

Methodology

Data Preparation for Toronto Data

At this stage, libraries useful for the data manipulation was imported. This includes: ProgressBar, BeautifulSoup, numpy, pandas, matplotlib, requests, Nominatim, matplotlib, KMeans # import k-means, folium and lxml.

1.
   i. We begin manipulation with our Toronto and New York data by converting it to dataframe.

[70]:

	PostalCode	Borough	Neighborhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

2. We then arranged the data by postal code

71]:

	PostalCode	Borough	Neighborhood
0	M1A	Not assigned	Not assigned
1	M1B	Scarborough	Malvern, Rouge
2	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
3	M1E	Scarborough	Guildwood, Morningside, West Hill
4	M1G	Scarborough	Woburn

3. The table shows that there are missing values “not assigned” hence the entire rows were removed

[72]:
```
df.drop(df[df['Borough']=="Not assigned"].index,axis=0, inplace=True)
df.head()
```

[72]:

	PostalCode	Borough	Neighborhood
1	M1B	Scarborough	Malvern, Rouge
2	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
3	M1E	Scarborough	Guildwood, Morningside, West Hill
4	M1G	Scarborough	Woburn
5	M1H	Scarborough	Cedarbrae

5. We now include gps coordinates from the table http://cocl.us/Geospatial_data

74]:
```
df_toronto = pd.merge(df, df_geo_coor, how='left', left_on = 'PostalCode', right_on = 'Postal Code')
# remove the "Postal Code" column
df_toronto.drop("Postal Code", axis=1, inplace=True)
df_toronto.head()
```

74]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	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

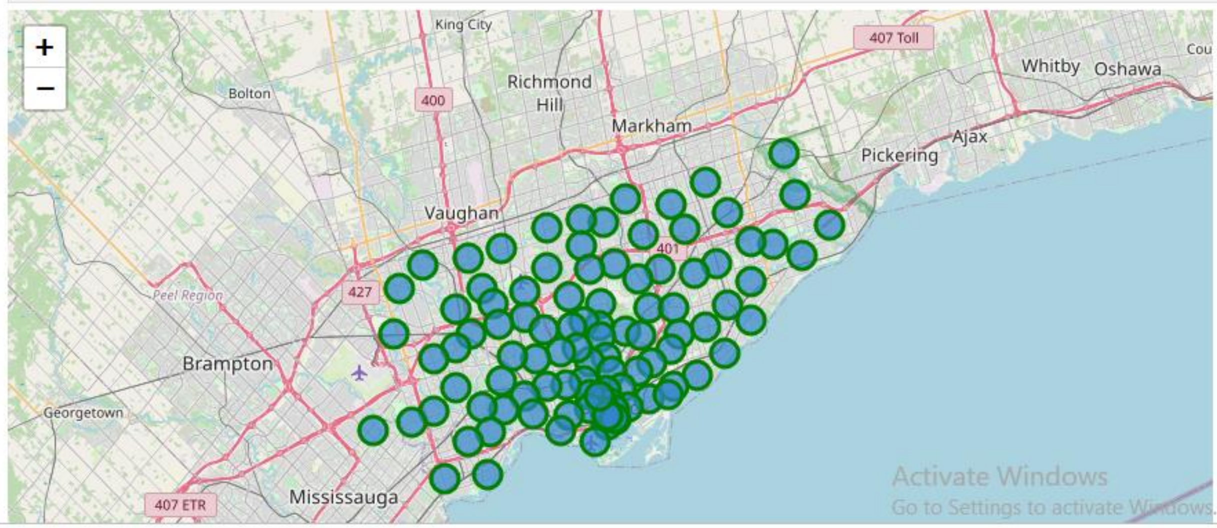
6. At this stage, the gps coordinate was merged with the dataframe and postal code column was dropped

```
df_toronto = pd.merge(df, df_geo_coor, how='left', left_on = 'PostalCode', right_on = 'Postal Code')
# remove the "Postal Code" column
df_toronto.drop("PostalCode", axis=1, inplace=True)
df_toronto.head()
```

	Borough	Neighborhood	Postal Code	Latitude	Longitude
0	Scarborough	Malvern, Rouge	M1B	43.806686	-79.194353
1	Scarborough	Rouge Hill, Port Union, Highland Creek	M1C	43.784535	-79.160497
2	Scarborough	Guildwood, Morningside, West Hill	M1E	43.763573	-79.188711
3	Scarborough	Woburn	M1G	43.770992	-79.216917
4	Scarborough	Cedarbrae	M1H	43.773136	-79.239476

7. Nominatim which is a geolocator was used to attached coordinates to locations on a grapgh and then graph was plotted using folium

```
map_toronto
```



8. Foursquare is used at this stage for analysis

```
[77]: #Foursquare Credentials
CLIENT_ID = 'D03ERUWRGPTNEAFVFNKLAT5YJVRQ1WMDRSQDPB2OZJRHHKE2' # your Foursquare ID
CLIENT_SECRET = 'ODJZVJTBPPWBMONE1RN1STL3ARS4KZJLBIRGYVJRZ0YIY3EC' # your Foursquare Secret
VERSION = '20180605' # Foursquare API version

print('Your credentails:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
```

```
Your credentails:
CLIENT_ID: D03ERUWRGPTNEAFVFNKLAT5YJVRQ1WMDRSQDPB2OZJRHHKE2
CLIENT_SECRET:ODJZVJTBPPWBMONE1RN1STL3ARS4KZJLBIRGYVJRZ0YIY3EC
```

9. We download all venues at a radius of 2000 around Malvern, Rouge and put it in a dataframe for manipulation.

34]:

	name	categories	lat	lng
0	African Rainforest Pavilion	Zoo Exhibit	43.817725	-79.183433
1	Images Salon & Spa	Spa	43.802283	-79.198565
2	Toronto Pan Am Sports Centre	Athletics & Sports	43.790623	-79.193869
3	Toronto Zoo	Zoo	43.820582	-79.181551
4	Polar Bear Exhibit	Zoo	43.823372	-79.185145

Data Modelling for Toronto

At this stage, we want to give meaningful insight to data by manipulation and analysis. I want to reveal patterns and structure within the data and hence get some insights from it.

1. Get list of venues around Toronto

```
89]: toronto_venues = getNearbyVenues(names=df_toronto['Neighborhood'],
                                   latitudes=df_toronto['Latitude'],
                                   longitudes=df_toronto['Longitude']
                                  )
```

```
Malvern, Rouge
Rouge Hill, Port Union, Highland Creek
Guildwood, Morningside, West Hill
Woburn
Cedarbrae
Scarborough Village
Kennedy Park, Ionview, East Birchmount Park
Golden Mile, Clairlea, Oakridge
Cliffside, Cliffcrest, Scarborough Village West
Birch Cliff, Cliffside West
Dorset Park, Wexford Heights, Scarborough Town Centre
Wexford, Maryvale
Agincourt
Clarks Corners, Tam O'Shanter, Sullivan
Milliken, Agincourt North, Steeles East, L'Amoreaux East
Steeles West, L'Amoreaux West
Upper Rouge
Hillcrest Village
Fairview, Henry Farm, Oriole
Bayview Village
York Mills. Silver Hills
```

2. Group by neighborhoods in Toronto

[90]:

```
toronto_venues.groupby('Neighborhood').count()
```

[90]:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Agincourt	4	4	4	4	4	4
Alderwood, Long Branch	9	9	9	9	9	9
Bathurst Manor, Wilson Heights, Downsview North	19	19	19	19	19	19
Bayview Village	4	4	4	4	4	4
Bedford Park, Lawrence Manor East	25	25	25	25	25	25
...	...	...	...	...	...	...
Willowdale, Willowdale East	33	33	33	33	33	33
Willowdale, Willowdale West	6	6	6	6	6	6
Woburn	5	5	5	5	5	5
Woodbine Heights	7	7	7	7	7	7
York Mills West	3	3	3	3	3	3

3. Analyse each neighborhoods in Toronto to get the frequency of visits to the various location.

1]:

```
# one hot encoding
toronto_onehot = pd.get_dummies(toronto_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
toronto_onehot['Neighborhood'] = toronto_venues['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [toronto_onehot.columns[-1]] + list(toronto_onehot.columns[:-1])
toronto_onehot = toronto_onehot[fixed_columns]

toronto_onehot.head()
```

1]:

	Yoga Studio	Accessories Store	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	...	Train Station	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnai Restau
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	

```
[92]: toronto_grouped = toronto_onehot.groupby('Neighborhood').mean().reset_index()
toronto_grouped
```

[92]:

	Neighborhood	Yoga Studio	Accessories Store	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...	Train Station	Vegetarian / Vegan Restaurant	Video Game Store	Vide Stor
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	...	0.0	0.0	0.0	0.00000
1	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	...	0.0	0.0	0.0	0.00000
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	...	0.0	0.0	0.0	0.00000
3	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	...	0.0	0.0	0.0	0.00000
4	Bedford Park, Lawrence	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.04	...	0.0	0.0	0.0	0.00000

4. Rank the venues according to how busy they are.

[27]:

	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
0	Agincourt	Lounge	Latin American Restaurant	Skating Rink	Breakfast Spot	Dog Run	Dessert Shop	Dim Sum Restaurant	Diner
1	Alderwood, Long Branch	Pizza Place	Gym	Coffee Shop	Dance Studio	Pharmacy	Skating Rink	Sandwich Place	Pub
2	Bathurst Manor, Wilson Heights, Downsview North	Bank	Coffee Shop	Mobile Phone Shop	Sandwich Place	Middle Eastern Restaurant	Supermarket	Restaurant	Ice Cream Shop
3	Bayview Village	Café	Bank	Japanese Restaurant	Chinese Restaurant	Women's Store	Department Store	Dim Sum Restaurant	Diner
4	Bedford Park, Lawrence Manor East	Coffee Shop	Sandwich Place	Italian Restaurant	Thai Restaurant	Restaurant	Pizza Place	Pharmacy	Café
5	Berczy Park	Coffee Shop	Cocktail Bar	Farmers Market	Café	Bakery	Restaurant	Beer Bar	Seafood Restaurant
6	Birch Cliff, Cliffside West	Café	College Stadium	Skating Rink	General Entertainment	Distribution Center	Department Store	Dessert Shop	Dim Sum Restaurant

Results for Toronto

Neighborhoods having venues with very busy Pharmacy outlets can be recommended as good location for situating a pharmacy outlet as that means a large market share.

```
neighborhoods_venues_sorted[neighborhoods_venues_sorted['1st Most Common Venue'] == 'Pharmacy']
```

	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
16	Clarks Corners, Tam O'Shanter, Sullivan	Pharmacy	Pizza Place	Fried Chicken Joint	Fast Food Restaurant	Italian Restaurant	Thai Restaurant	Chinese Restaurant	Gas Station	Convenience Store	Noodle House
27	Eringate, Bloordale Gardens, Old Burnhamthorpe...	Pharmacy	Beer Store	Pet Store	Pizza Place	Coffee Shop	Café	Liquor Store	General Entertainment	Donut Shop	Dog Run
91	Willowdale, Willowdale West	Pharmacy	Discount Store	Grocery Store	Pizza Place	Coffee Shop	Bank	Department Store	Dessert Shop	Dim Sum Restaurant	Diner
93	Woodbine Heights	Pharmacy	Video Store	Park	Beer Store	Skating Rink	Curling Ice	Athletics & Sports	Comic Shop	Concert Hall	College Rec Center

NEW YORK CITY

Following similar data analysis steps for Brooklyn, Newyork

1. Tranform the data into a pandas dataframe

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

2. Nominatim which is a geolocator was used to attached coordinates to locations on a grapgh and then graph was plotted using folium

```
1]: address = 'New York City, NY'

geolocator = Nominatim(user_agent="ny_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of New York City are {}, {}.'.format(latitude, longitude))
```

```
The geograpical coordinate of New York City are 40.7127281, -74.0060152.
```

3. Let's slice the original dataframe and create a new dataframe of the Brooklyn data.

```
114]: brooklyn_data = neighborhoods[neighborhoods['Borough'] == 'Brooklyn'].reset_index(drop=True)
brooklyn_data.head()
```

114]:

	Borough	Neighborhood	Latitude	Longitude
0	Brooklyn	Bay Ridge	40.625801	-74.030621
1	Brooklyn	Bensonhurst	40.611009	-73.995180
2	Brooklyn	Sunset Park	40.645103	-74.010316
3	Brooklyn	Greenpoint	40.730201	-73.954241
4	Brooklyn	Gravesend	40.595260	-73.973471

4. Using Foursquare to get venues around the neighborhood

```
: # type your answer here
#The correct answer is:
LIMIT = 100 # limit of number of venues returned by Foursquare API



radius = 2000 # define radius



# create URL
url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION,
    neighborhood_lat,
    neighborhood_long,
    radius,
    LIMIT)
url # display URL
```


5. Get the Categories of the Venue

This is separate from the ipykernel package so we can avoid doing imports until

[125]:

	name	categories	lat	lng
0	Pilo Arts Day Spa and Salon	Spa	40.624748	-74.030591
1	Bagel Boy	Bagel Shop	40.627896	-74.029335
2	Ho' Brah Taco Joint	Taco Place	40.622960	-74.031371
3	Pegasus Cafe	Breakfast Spot	40.623168	-74.031186
4	Karam	Middle Eastern Restaurant	40.622931	-74.028316

6. Explore Neighborhoods in Brooklyn

.28]:

```
brooklyn_venues = getNearbyVenues(names=brooklyn_data['Neighborhood'],
                                   latitudes=brooklyn_data['Latitude'],
                                   longitudes=brooklyn_data['Longitude']
                                  )
```

```
Bay Ridge
Bensonhurst
Sunset Park
Greenpoint
Gravesend
Brighton Beach
Sheepshead Bay
Manhattan Terrace
Flatbush
Crown Heights
East Flatbush
Kensington
Windsor Terrace
Prospect Heights
Brownsville
Williamsburg
Bushwick
Bedford Stuyvesant
```

7. Analyse the frequency of patronage of various venues

```
brooklyn_venues.groupby('Neighborhood').count()
```

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Bath Beach	52	52	52	52	52	52
Bay Ridge	84	84	84	84	84	84
Bedford Stuyvesant	27	27	27	27	27	27
Bensonhurst	34	34	34	34	34	34
Bergen Beach	5	5	5	5	5	5
...	...	...	...	...	...	...
Vinegar Hill	28	28	28	28	28	28
Weeksville	17	17	17	17	17	17
Williamsburg	36	36	36	36	36	36
Windsor Terrace	27	27	27	27	27	27
Wingate	20	20	20	20	20	20

70 rows × 6 columns

8. Analyze each Brooklyn Neighborhood to show the frequency of use

	Neighborhood	Yoga Studio	Accessories Store	Adult Boutique	Airport Terminal	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	...	Vegetarian / Vegan Restaurant	Video Game Store
0	Bath Beach	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.000000	...	0.0	0.019231
1	Bay Ridge	0.000000	0.0	0.0	0.0	0.035714	0.000000	0.0	0.0	0.000000	...	0.0	0.011905
2	Bedford Stuyvesant	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.000000	...	0.0	0.000000
3	Bensonhurst	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.000000	...	0.0	0.000000
4	Bergen Beach	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.000000	...	0.0	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...
65	Vinegar Hill	0.000000	0.0	0.0	0.0	0.035714	0.035714	0.0	0.0	0.071429	...	0.0	0.000000
66	Weeksville	0.000000	0.0	0.0	0.0	0.058824	0.000000	0.0	0.0	0.000000	...	0.0	0.000000
67	Williamsburg	0.027778	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.027778	...	0.0	0.000000
68	Windsor Terrace	0.000000	0.0	0.0	0.0	0.037037	0.037037	0.0	0.0	0.000000	...	0.0	0.000000
69	Wingate	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.000000	...	0.0	0.000000

9. Get top ten most commonly used venues

	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	Bath Beach	Chinese Restaurant	Pharmacy	Pizza Place	Gas Station	Bubble Tea Shop	Donut Shop	Fast Food Restaurant	Italian Restaurant	Cantonese Restaurant	Spanish Restaurant
1	Bay Ridge	Italian Restaurant	Spa	Pizza Place	Greek Restaurant	American Restaurant	Bar	Mediterranean Restaurant	Sandwich Place	Chinese Restaurant	Playground
2	Bedford Stuyvesant	Coffee Shop	Pizza Place	Café	Bar	Tiki Bar	Bus Stop	Fried Chicken Joint	Boutique	New American Restaurant	Gift Shop
3	Bensonhurst	Chinese Restaurant	Flower Shop	Italian Restaurant	Sushi Restaurant	Grocery Store	Donut Shop	Bakery	Ice Cream Shop	Cha Chaan Teng	Bank
4	Bergen Beach	Harbor / Marina	Baseball Field	Playground	Athletics & Sports	Filipino Restaurant	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Field
5	Boerum Hill	Dance Studio	Coffee Shop	Bar	Furniture / Home Store	French Restaurant	Arts & Crafts Store	Sandwich Place	Bakery	Cocktail Bar	Grocery Store
6	Borough Park	Bank	Café	Deli / Bodega	Pizza Place	Fast Food Restaurant	Pharmacy	Restaurant	Chinese Restaurant	Bistro	Bakery

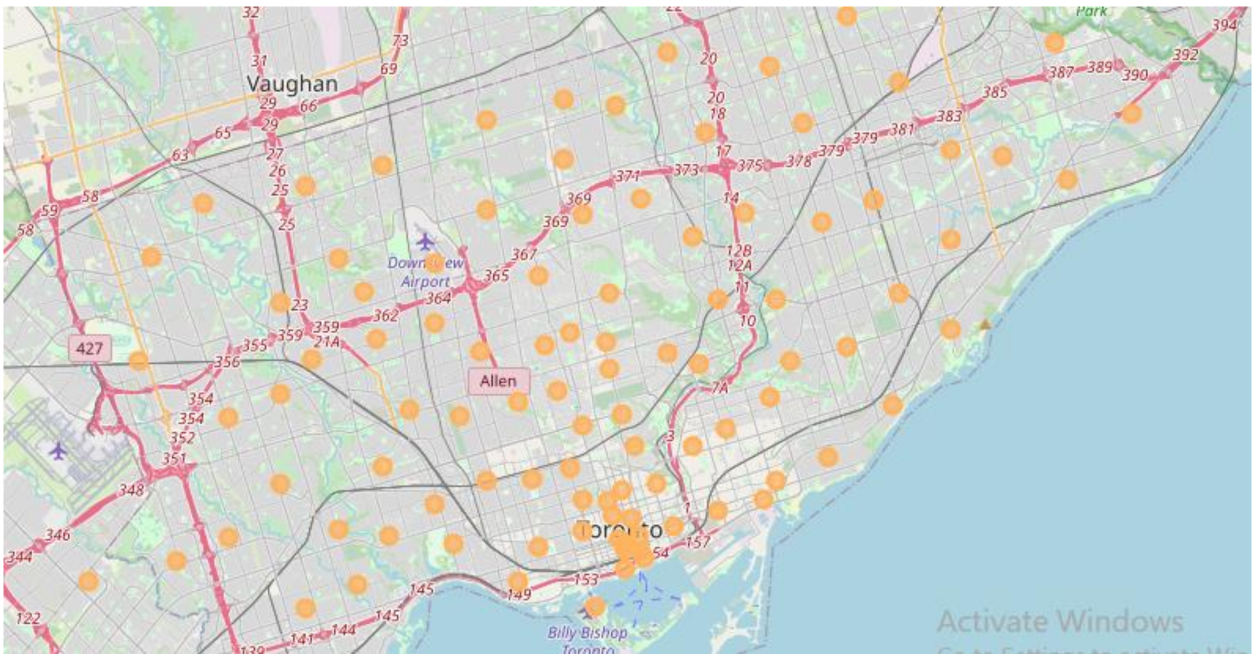
Results for Brooklyn New York

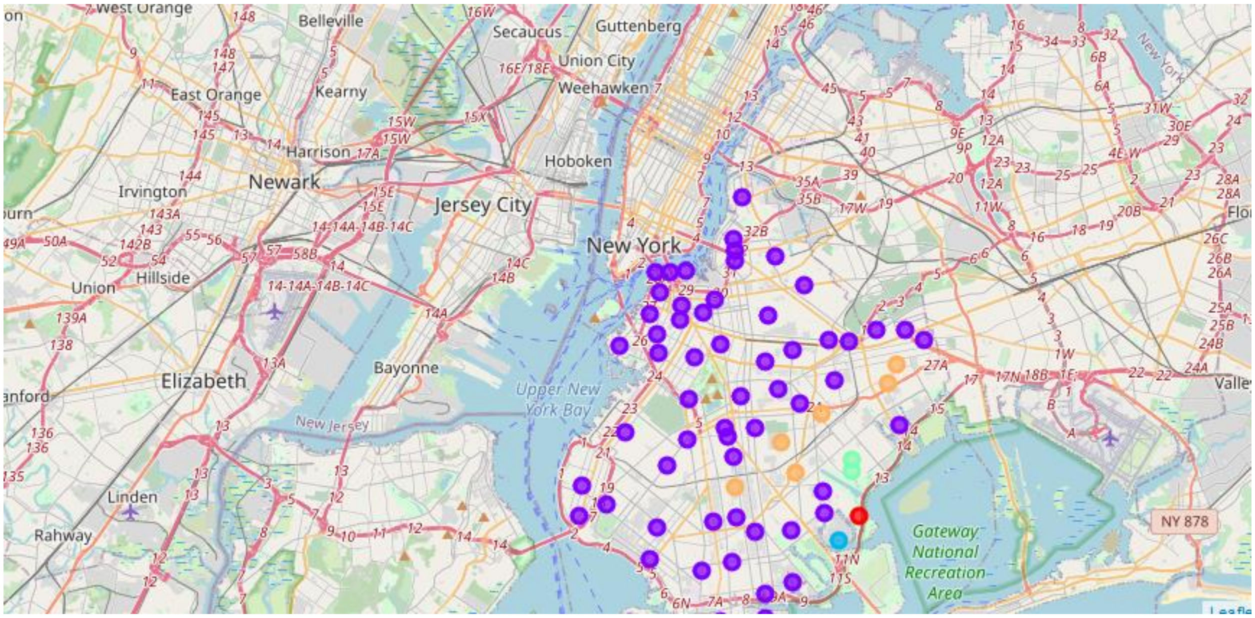
Neighborhoods having venues with very busy Pharmacy outlets can be recommended as good location for situating a pharmacy outlet as that means a large market share

	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
29	Flatlands	Pharmacy	Fried Chicken Joint	Fast Food Restaurant	Caribbean Restaurant	Deli / Bodega	Park	Bus Station	Nightclub	Lounge	Chinese Restaurant

EVALUATION

Comparing between Toronto and New York using K Clusters





Discussion

From the above analysis, results and evaluation, I will make the following recommendation to my client:

1. Between Newyork and Toronto, with respect to serenity of the neighborhoods and relative calm, I will suggest they emigrate to Toronto Canada
2. Within the Toronto neighborhood, it will be better to situate a pharmacy in an already where there is high frequency to Pharmacy outlets as that signifies a large market share and good competition

	Neighborhood	1st Most Common Venue
16	Clarks Corners, Tam O'Shanter, Sullivan	Pharmacy
27	Eringate, Bloordale Gardens, Old Burnhamthorpe...	Pharmacy
91	Willowdale, Willowdale West	Pharmacy
93	Woodbine Heights	Pharmacy

Conclusion

We can safely conclude that with the help of data science, strategic and business decisions can be more accurate.