The Battle of Neighborhoods – Toronto Data Analysis

Introduction

Background:

The purpose of this project is to help the people wanting to immigrate to major cities of the world with research on good neighborhoods and ultimately make a smart decision.

Essentially this project will provide a detailed information of the neighborhoods in a city including the various catergorical venues including yoga centers, coffee shops, restaurants etc.

This project also aims at providing the list of secondary schools and other information on schools like pass percentage, population etc.

Assuming that an immigrant wants to setup a coffee shop business, this project will be providing with a list of neighborhoods with least number of 'coffee shops' for someone to setup as a business.

In totality it will help people get an awareness on a new city. It will provide a comparative analysis of various neighbor hoods for settlement and opening up a particular business.

Problem Statement:

The purpose of this project is to provide information on:

- Good neighborhoods with accessible venues
- Good secondary school location and information
- Good neighborhood to setup a Coffee Shop

Location to be analyzed:

Toronto, CA is a very popular destination for immigration for staying and setting up business. Therefore, for this project the Toronto city, its neighborhoods, school districts and various categorical venues will be used.

K-Means clustering:

The similarities or dissimilarities between two neighborhoods in a city could be visualized by segmenting them into various clusters utilizing the k-means clustering machine learning algorithm. So, this project aims at clustering and making sense of data obtained from this clustering technique

Location Data Analysis - Four Square API:

Forsquare is one of the location data providers. A developer account has been created and credentials have been obtained. Due to a limited number of accesses for this developer account, there will be some restrictions on the radius and count of venues search.

Data Analysis and Libraries:

Along with the Foursquare API, Python and its associated libraries and packages will be used for data analysis and visualization.

Pandas: Dataframe creation and manipulation

Numpy: Mathematical Analysis

Matplotlib: Python plotting module

Folium: Interactive leaflet map creation

Scikit Learn: Implementing k-means clustering algorithm

JSON: Handling JSON files

Geocoder: Retrieving location data

Beautiful Soup and Requests: Web scrapping and data acquisition

Data Section

Toronto Neighborhood Data:

The Toronto neighborhood data is obtained form:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The Toronto secondary school data is obtained from:

https://en.wikipedia.org/wiki/List_of_secondary_schools_in_the_Toronto_District_School_Board

Also, the data set with Toronto neighborhood spatial data used in Week 3 Toronto data analysis is also be used. https://github.com/karthikeyachalla1992/Toronto-Data-Project/blob/master/GC.csv

Data obtained from Four Square API:

Four Square us a location data provider with information on venues nearby a neighborhood. It provides:

- Neighborhood
- Neighborhood Latitude
- Neighborhood Longitude
- Venue
- Name of the Venue
- Venue latitude
- Venue longitude
- Venue Category

Determining the location to live

The immigrants look for better schools in the areas they want to settle. In this part of the project, the neighborhood and secondary school data has been acquired and formatted to form a table.

This table shows the school name, borough, population and pass percentage information.

	Name	Borough	Math Score	ESL population	OSSLT Pass Percentage	Postal code
0	A. Y. Jackson Secondary School	North York	93%	84%	90%	M2H
1	Agincourt Collegiate Institute	Scarborough	90%	79%	92%	M1S
2	Albert Campbell Collegiate Institute	Scarborough	84%	86%	80%	M1V
3	Birchmount Park Collegiate Institute	Scarborough	67%	43%	70%	M1N
4	Bloor Collegiate Institute	Toronto	91%	65%	91%	M6H

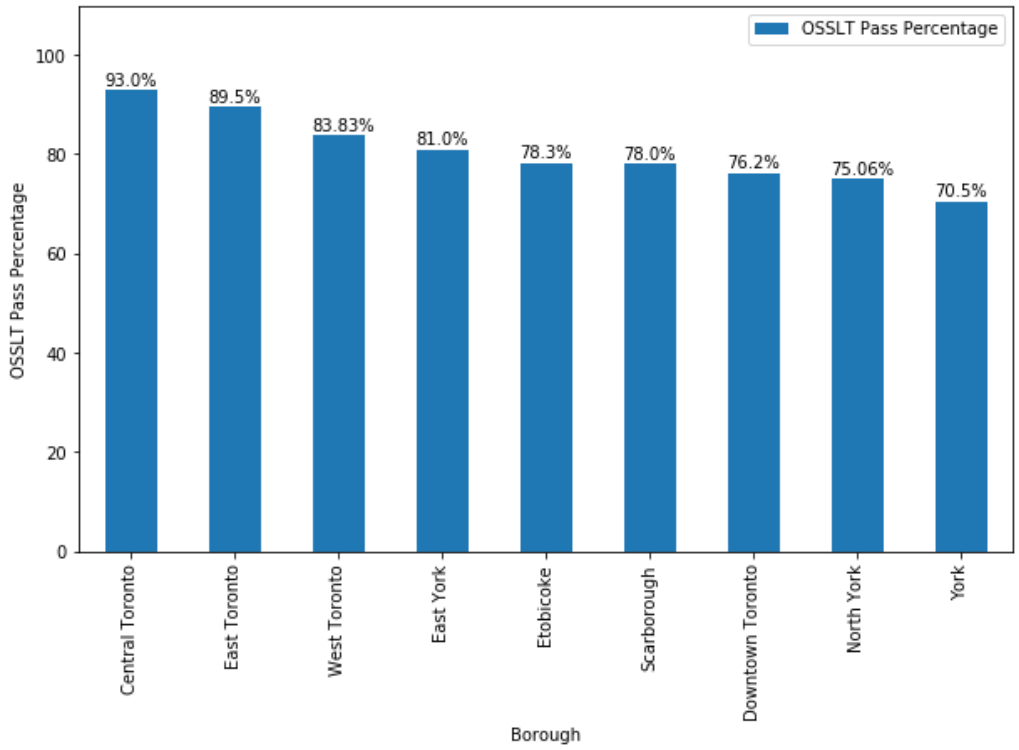
Then the neighborhood data of Toronto is extracted to understand to what boroughs each neighborhood belongs based on postal codes.

	Postal code	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park / Harbourfront
3	M6A	North York	Lawrence Manor / Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park / Ontario Provincial Government

Then the above two tables are merged to get the neighborhood information for each school. Only OSSLT Pass Percentage is used as a metric to evaluate the school's performance.

	Postal code	Borough	Neighborhood	Name	OSSLT Pass Percentage
0	M3A	North York	Parkwoods	George S. Henry Academy	68
1	M3A	North York	Parkwoods	Victoria Park Collegiate Institute	84
2	M6A	North York	Lawrence Manor / Lawrence Heights	John Polanyi Collegiate Institute	76
3	M9A	Etobicoke	Islington Avenue	Etobicoke Collegiate Institute	84
4	M9A	Etobicoke	Islington Avenue	Richview Collegiate Institute	94

Then a bar graph is created to provide the mean pass percentage of the top boroughs to determine the top neighborhood to settle.



The top 5 schools from top 3 boroughs are displayed in the table below.

	Postal code	Borough	Neighborhood	Name	OSSLT Pass Percentage
0	M6P	West Toronto	High Park / The Junction South	Ursula Franklin Academy	99
1	M4P	Central Toronto	Davisville North	North Toronto Collegiate Institute	98
2	M4R	Central Toronto	North Toronto West	Lawrence Park Collegiate Institute	96
3	M4E	East Toronto	The Beaches	Malvern Collegiate Institute	94
4	M6P	West Toronto	High Park / The Junction South	Humberside Collegiate Institute	94

Based on the above data analysis the best neighborhood to settle is determined.

```
bestneighborhoods = tds3['Neighborhood'].value_counts().to_frame()
bestneighborhoods.rename(columns={'Neighborhood':'No. of Schools'},inplace=True)
bestneighborhoods.index.name='Neighborhood'
bestneighborhoods.reset_index(inplace=True)
bestneighborhoods.sort_values('No. of Schools', inplace=True)
bestneighborhood= bestneighborhoods['Neighborhood'][0]


print('The best neighborhood to live according to secondary school pass percentages is {}'.format(bestneighborhood))
```

```
The best neighborhood to live according to secondary school pass percentages is High Park / The Junction South
```

Determining the best neighborhood to open coffee shop business

The decision is made on the basis that the coffee shop is not in top 5 venues for any particular neighborhood.

The CSV file is used to extract the neighborhood information along with latitude and longitude information.

	Borough	Neighborhood	Latitude	Longitude
0	North York	Parkwoods	43.753259	-79.329656
1	North York	Victoria Village	43.725882	-79.315572
2	Downtown Toronto	Regent Park / Harbourfront	43.654260	-79.360636
3	North York	Lawrence Manor / Lawrence Heights	43.718518	-79.464763
4	Downtown Toronto	Queen's Park / Ontario Provincial Government	43.662301	-79.389494

A map is created to show all the neighborhoods in Toronto.

Foursquare API is used to get all the information of venues and locations around the neighborhoods.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
4	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

Venue category is an important column for us to determine the neighborhoods without coffee shop as their top venues.

Then, a table is created with all the neighborhoods along with their top 5 venues.

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Agincourt	Breakfast Spot	Lounge	Latin American Restaurant	Skating Rink	Massage Studio
1	Alderwood / Long Branch	Pizza Place	Coffee Shop	Pharmacy	Sandwich Place	Gym
2	Bathurst Manor / Wilson Heights / Downsview North	Bank	Coffee Shop	Pizza Place	Pharmacy	Deli / Bodega
3	Bayview Village	Bank	Chinese Restaurant	Japanese Restaurant	Café	Accessories Store
4	Bedford Park / Lawrence Manor East	Sandwich Place	Restaurant	Coffee Shop	Italian Restaurant	Japanese Restaurant

Then all the neighborhoods with coffee shop in their top 5 venues are removed

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Agincourt	Breakfast Spot	Lounge	Latin American Restaurant	Skating Rink	Massage Studio
1	Bayview Village	Bank	Chinese Restaurant	Japanese Restaurant	Café	Accessories Store
2	Birch Cliff / Cliffside West	General Entertainment	College Stadium	Skating Rink	Café	Accessories Store
3	Business reply mail Processing CentrE	Yoga Studio	Garden Center	Pizza Place	Comic Shop	Restaurant
4	CN Tower / King and Spadina / Railway Lands / ...	Airport Service	Airport Lounge	Plane	Boutique	Airport Food Court

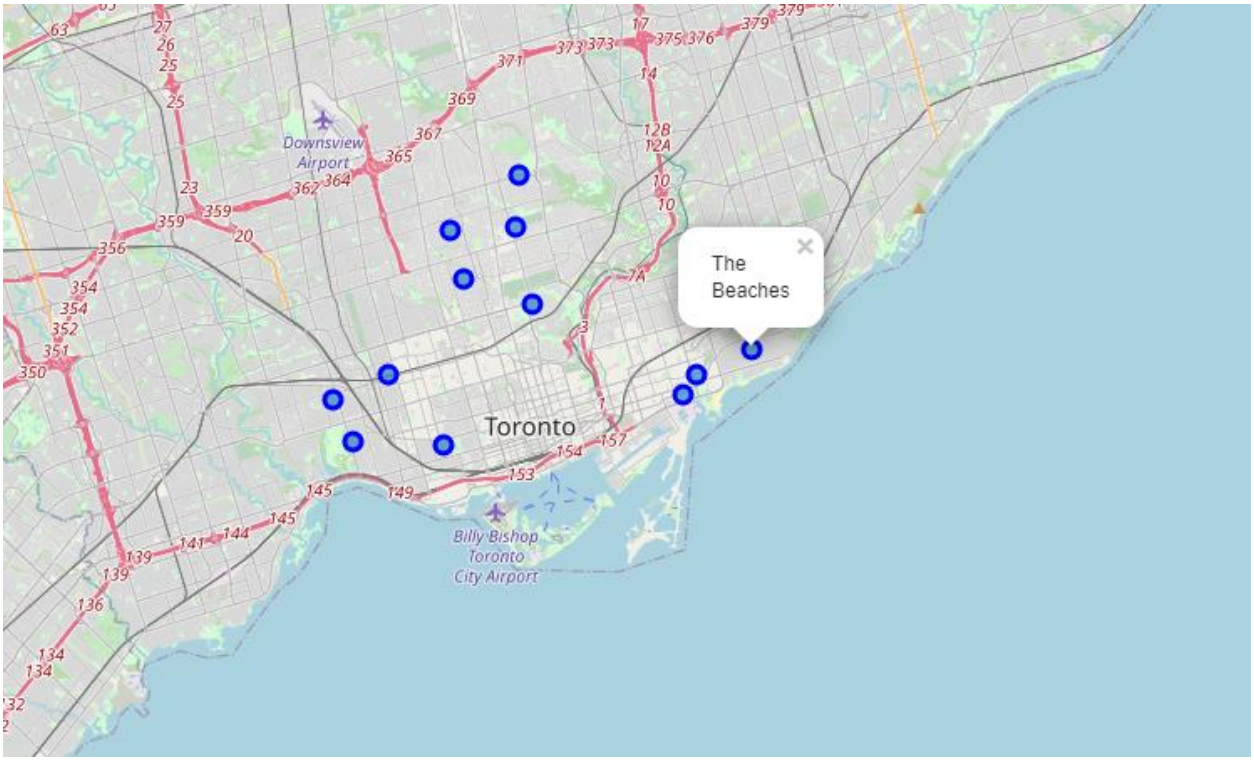
From the data analysis of the schools information it has been determined that the top 3 boroughs to settle would be West, East and Central Toronto.

Therefore, only the neighborhoods in these 3 boroughs without a coffee shop in their top 5 venues will be analyzed.

Then a table has been created to show the best neighborhoods to setup a coffee shop business.

	Borough	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	East Toronto	The Beaches	Health Food Store	Pub	Neighborhood	Trail	Monument / Landmark
1	West Toronto	Dufferin / Dovercourt Village	Pharmacy	Bakery	Brewery	Grocery Store	Middle Eastern Restaurant
2	West Toronto	Little Portugal / Trinity	Bar	Restaurant	Asian Restaurant	Café	Vietnamese Restaurant
3	East Toronto	India Bazaar / The Beaches West	Park	Sandwich Place	Fast Food Restaurant	Italian Restaurant	Pet Store
4	Central Toronto	Lawrence Park	Park	Swim School	Bus Line	Accessories Store	Mexican Restaurant
5	Central Toronto	Roselawn	Garden	Home Service	Middle Eastern Restaurant	Moroccan Restaurant	Monument / Landmark
6	Central Toronto	Davisville North	Breakfast Spot	Department Store	Park	Sandwich Place	Food & Drink Shop
7	Central Toronto	Forest Hill North & West	Bus Line	Sushi Restaurant	Jewelry Store	Trail	Accessories Store
8	West Toronto	High Park / The Junction South	Mexican Restaurant	Café	Thai Restaurant	Arts & Crafts Store	Fried Chicken Joint
9	West Toronto	Parkdale / Roncesvalles	Gift Shop	Movie Theater	Cuban Restaurant	Bar	Breakfast Spot
10	Central Toronto	Moore Park / Summerhill East	Summer Camp	Playground	Mexican Restaurant	Monument / Landmark	Molecular Gastronomy Restaurant
11	East Toronto	Business reply mail Processing CentrE	Yoga Studio	Garden Center	Pizza Place	Comic Shop	Restaurant

A map has been created to visualize the above.



K-Means Clustering – Analyzing the Neighborhoods and Venues using Machine Learning

First, the cluster labels are created.

```
# clustering the neighborhoods

# set number of clusters
kclusters = 5

tcluster = tgroup.drop('Neighborhoods', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(tcluster)

# check cluster labels generated for each row in the dataframe
kmeans.labels_.astype(int)
#tsorted.drop('Cluster Labels',axis=1,inplace=True)
```

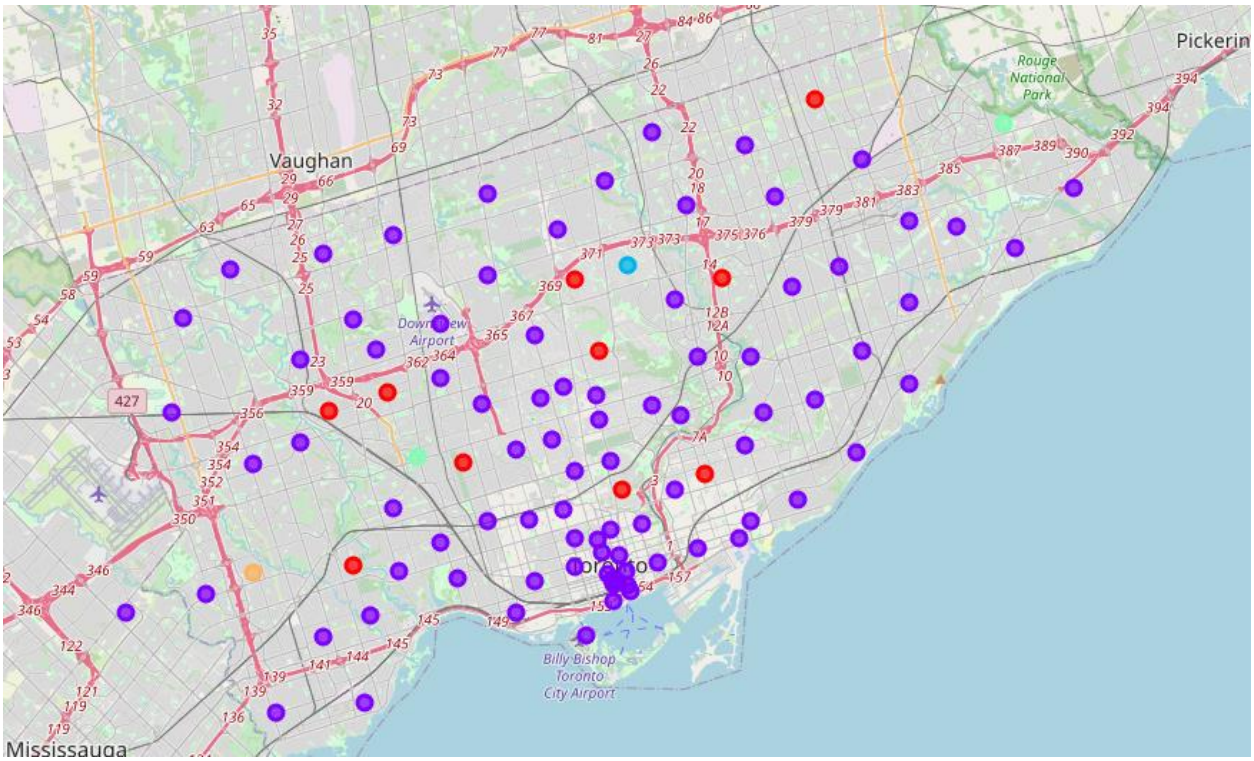
```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3,
       1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 0, 1, 1, 3, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 4, 1,
       0, 1, 1, 1, 1, 2, 0])
```

Cluster labels are added to the table with neighborhood and venues information.

	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
0	North York	Parkwoods	43.753259	-79.329656	0	Park	Food & Drink Shop	Accessories Store	Mexican Restaurant	Molecular Gastronomy Restaurant
1	North York	Victoria Village	43.725882	-79.315572	1	Pizza Place	Hockey Arena	Portuguese Restaurant	Coffee Shop	Middle Eastern Restaurant
2	Downtown Toronto	Regent Park / Harbourfront	43.654260	-79.360636	1	Coffee Shop	Bakery	Pub	Park	Breakfast Spot
3	North York	Lawrence Manor / Lawrence Heights	43.718518	-79.464763	1	Accessories Store	Miscellaneous Shop	Clothing Store	Gift Shop	Furniture / Home Store
4	Downtown Toronto	Queen's Park / Ontario Provincial Government	43.662301	-79.389494	1	Coffee Shop	Diner	Sushi Restaurant	Yoga Studio	Café

Based on this cluster label we can map them to visualize the location distinction.

Then we can analyze them label by label to understand how the algorithms has segregated the city data.



Each color dot represents one cluster, for this example only 5 clusters are used.

Analyzing each cluster separately.

Cluster 0 represents the neighborhoods with Park as their 1st most common venue.

```
tmerged.loc[tmerged['Cluster Labels'] == 0, tmerged.columns[[1] + list(range(5, tmerged.shape[1]))]]
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Parkwoods	Park	Food & Drink Shop	Accessories Store	Mexican Restaurant	Molecular Gastronomy Restaurant
20	Caledonia-Fairbanks	Park	Women's Store	Pool	Accessories Store	Mexican Restaurant
34	East Toronto	Pizza Place	Park	Convenience Store	Accessories Store	Middle Eastern Restaurant
50	North Park / Maple Leaf Park / Upwood Park	Bakery	Construction & Landscaping	Park	Miscellaneous Shop	Motel
60	Lawrence Park	Park	Swim School	Bus Line	Accessories Store	Mexican Restaurant
63	Weston	Park	Convenience Store	Accessories Store	Mexican Restaurant	Monument / Landmark
65	York Mills West	Park	Convenience Store	Bank	Accessories Store	Miscellaneous Shop
83	Milliken / Agincourt North / Steeles East / L'...	Park	Playground	Accessories Store	Mexican Restaurant	Monument / Landmark
89	Rosedale	Park	Trail	Playground	Accessories Store	Monument / Landmark
95	The Kingsway / Montgomery Road / Old Mill North	Park	River	Accessories Store	Middle Eastern Restaurant	Monument / Landmark

Cluster 2 represents the neighborhoods with Park as their 1st most common venue.

```
tmerged.loc[tmerged['Cluster Labels'] == 3, tmerged.columns[[1] + list(range(5, tmerged.shape[1]))]]
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
5	Malvern / Rouge	Fast Food Restaurant	Accessories Store	Mexican Restaurant	Monument / Landmark	Molecular Gastronomy Restaurant
55	Del Ray / Mount Dennis / Keelsdale and Silvert...	Fast Food Restaurant	Bar	Sandwich Place	Accessories Store	Middle Eastern Restaurant

Similarly, other clusters can be analyzed.

Conclusion

The Toronto neighborhood data and secondary school data retrieved from the wiki pages have been analyzed and presented in the form of tables and maps for an immigrant to choose the place to live and setup a coffee shop business.