PROGRAM: IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

COURSE: APPLIED DATA SCIENCE CAPSTONE

BATTLE OF NEIGHBORHOODS

Title: Leveraging Location Data to Determine a Suitable Borough/Neighborhood to Open a Baseball Store

By

Ugochukwu Paul Emegwoako

Contents

Introduction

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2,731,571 as of 2016 (Toronto, n.d.). It is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. Toronto is represented in five major league sports, with teams in hockey, baseball, basketball, football, and soccer. As a centre of business, getting a good location for any business type is of the utmost importance.

Aim

This project aims to leverage location data to determine a suitable borough/neighborhood to open a baseball store.

Importance

This project will be important to any business professional as this will give any kind of business advantageous insights on how likely opening a store in a neighborhood will be profitable based on location data alone. This can also be extended to any kind of business and its requirements by adding other necessary data.

Data

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

This was used to get the postal codes and names of the neighborhoods and boroughs in Toronto

- Geodata: https://cocl.us/Geospatial_data

This was used to get the latitude and longitude of each borough gotten from the Wikipedia page.

- Foursquare:

This was used to get the venues in each borough as well as the venue latitude, longitude and venue category, in each borough.

Methodology

I started by reading the Wikipedia html into a Pandas data frame and proceeded to clean up my data. The “Not assigned” values were then changed to “NAN”. After the html link was read directly into the data frame, I had multiple pages from the website. I selected the table needed which contained data pertaining to postal code, borough, and neighborhood. The “NAN” values were then dropped. I noticed that there might be a possible conflict in naming, so I changed “neighbourhood” to “neighborhood” to avoid any possible future errors.

After cleaning up the data, my table had 103 rows and 3 columns consisting of Postal code, borough, and neighborhood. The next step was to use the geo spatial data (https://cocl.us/Geospatial_data) as provided by Coursera. This data contained the latitude and longitude of the postal codes in Toronto. It was the combined with our “T_data”, resulting in our Geodata data frame shown below.

	Postal Code	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
...	...	...	...	...	...
98	M9N	York	Weston	43.706876	-79.518188
99	M9P	Etobicoke	Westmount	43.696319	-79.532242
100	M9R	Etobicoke	Kingsview Village, St. Phillips, Martin Grove ...	43.688905	-79.554724
101	M9V	Etobicoke	South Steeles, Silverstone, Humbergate, Jamest...	43.739416	-79.588437
102	M9W	Etobicoke	Northwest, West Humber - Clairville	43.706748	-79.594054

103 rows × 5 columns

Figure 1: Geodata containing Latitude and longitude values of each neighborhood

With this data and using folium maps, I was then able to generate a map of the city with the blue circles representing the different boroughs in the city of Toronto. See figure below.

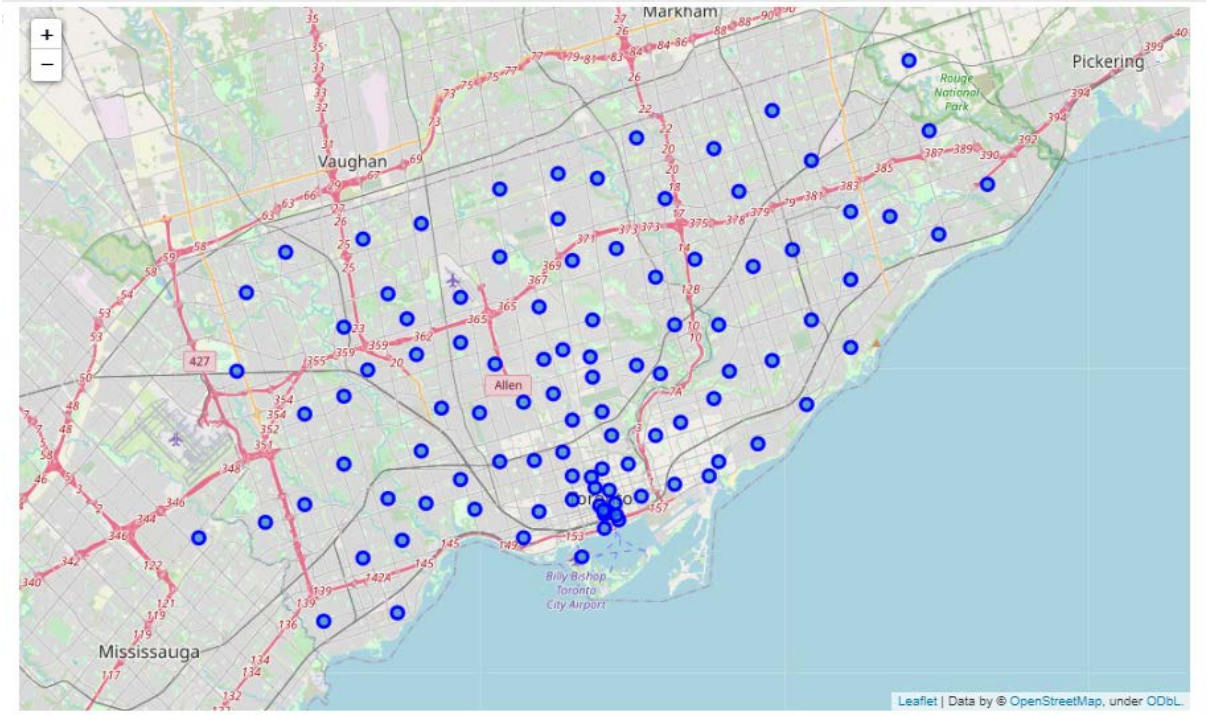

Figure 2: Map of Toronto and its neighborhoods

Next, I was able to use the data generated previously. This was done by passing each row into the foursquare api. This gave results consisting of venues, venue latitudes, venue longitude, etc. The figure below shows the first five results gotten from passing our data into a get request.

```
(2167, 7)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	SEBS Engineering Inc. (Sustainable Energy and ...	43.782371	-79.156820	Construction & Landscaping
3	Guildwood, Morningside, West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store

Figure 3:Tdata_venues showing neighborhood and venue category

A total of 2167 venues were gotten. I then cleaned the data and regrouped the venues. A total of 277 unique venue categories was gotten. I was then able to group all of the data into 1 table with the neighborhood included and each venue category frequency was then calculated. The table was then reordered to have each neighborhood and the top 10 locations/venue categories. The figure below shows the first four rows of my data.

	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	Lounge	Skating Rink	Latin American Restaurant	Breakfast Spot	Clothing Store	Donut Shop	Discount Store	Distribution Center	Dive Bar	Dog Run
1	Alderwood, Long Branch	Pizza Place	Pharmacy	Gym	Coffee Shop	Sandwich Place	Pub	Distribution Center	Dessert Shop	Dim Sum Restaurant	Diner
2	Bathurst Manor, Wilson Heights, Downsview North	Coffee Shop	Bank	Chinese Restaurant	Bridal Shop	Sandwich Place	Diner	Restaurant	Deli / Bodega	Middle Eastern Restaurant	Supermarket
3	Bayview Village	Café	Japanese Restaurant	Bank	Chinese Restaurant	Diner	Distribution Center	Dive Bar	Dog Run	Doner Restaurant	Donut Shop
4	Bedford Park, Lawrence Manor East	Coffee Shop	Sandwich Place	Italian Restaurant	Greek Restaurant	Sushi Restaurant	Pharmacy	Pizza Place	Pub	Café	Restaurant

Figure 4;Data showing neighborhoods and the top 10 venues

Kmeans clustering was used in this project. Before clustering, the optimal number of clusters needed to be determined. To determine this, I used the Silhouette score. Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. This resulted in the optimal number of clusters of 6. The figure below shows a graph of kmeans cluster number vs silhouette scores.

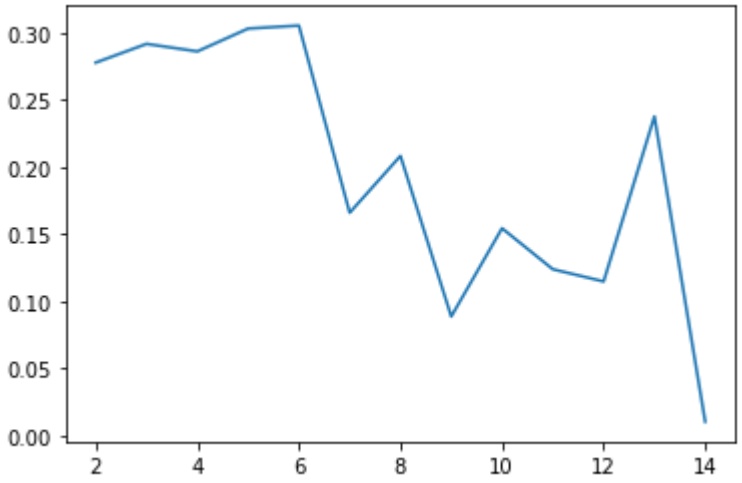

Figure 5: Graph of number of clusters vs silhouette score

Setting the number of clusters as 6, I was then able to use the Kmeans clustering algorithm on my cleaned data set and plotted this on the map like before.

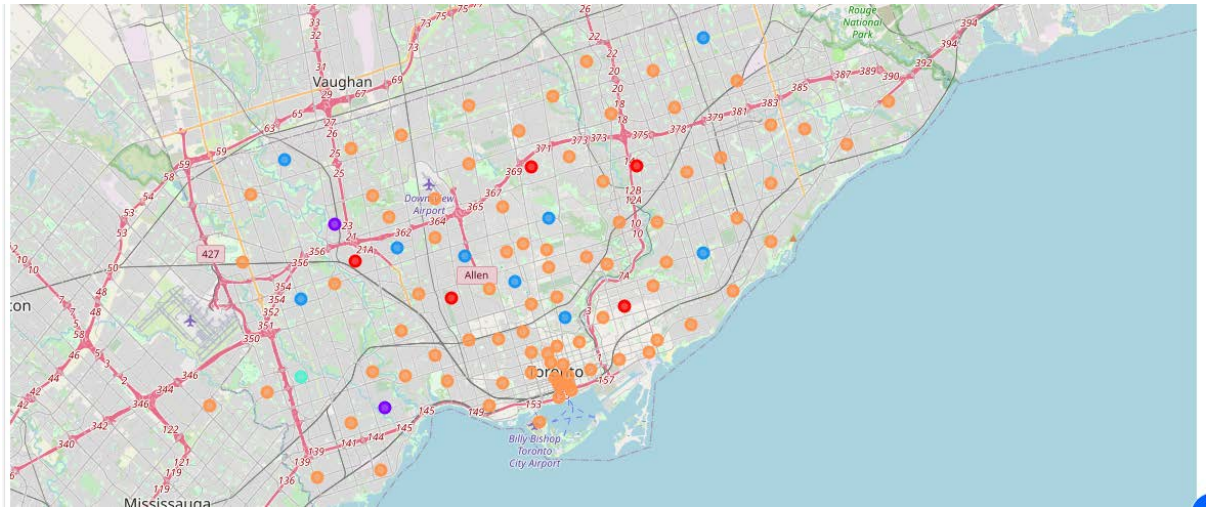

Figure 6:Clustered map of Toronto and its neighborhoods

Results

Figures 7 to 12 shows the six different clusters labeled A to F.

	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
7	Scarborough	0.0	Bus Line	Bakery	Soccer Field	Ice Cream Shop	Bus Station	Metro Station	Intersection	Park	German Restaurant	General Travel
14	Scarborough	0.0	Park	Playground	Bakery	Intersection	Women's Store	Doner Restaurant	Diner	Discount Store	Distribution Center	Dive Bar
44	Central Toronto	0.0	Park	Swim School	Bus Line	Women's Store	Doner Restaurant	Discount Store	Distribution Center	Dive Bar	Dog Run	Donut Shop
50	Downtown Toronto	0.0	Park	Playground	Tennis Court	Trail	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Donut Shop	Department Store
64	Central Toronto	0.0	Trail	Jewelry Store	Mexican Restaurant	Sushi Restaurant	Park	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Donut Shop
72	North York	0.0	Park	Pizza Place	Bakery	Japanese Restaurant	Pub	Dog Run	Diner	Discount Store	Distribution Center	Dive Bar
79	North York	0.0	Park	Bakery	Construction & Landscaping	Women's Store	Drugstore	Distribution Center	Dive Bar	Dog Run	Doner Restaurant	Donut Shop
96	North York	0.0	Pizza Place	Furniture / Home Store	Ethiopian Restaurant	Escape Room	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Dessert Shop	Doner Restaurant
100	Etobicoke	0.0	Pizza Place	Park	Bus Line	Sandwich Place	Dog Run	Diner	Discount Store	Distribution Center	Dive Bar	Doner Restaurant

Figure 7: CLUSTER #A

Cluster #A above has parks, pizza place and bakery as its most frequent venue.

	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
94	Etobicoke	1.0	Print Shop	Donut Shop	Diner	Discount Store	Distribution Center	Dive Bar	Dog Run	Doner Restaurant	Women's Store	Dessert Shop

Figure 8: CLUSTER #B

Cluster #B above has print shop as its most frequent venue.

	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Scarborough	2.0	Fast Food Restaurant	Donut Shop	Diner	Discount Store	Distribution Center	Dive Bar	Dog Run	Doner Restaurant	Drugstore	Farmers Market

Figure 9: CLUSTER #C

Cluster #C above has fast food as its most common venue. It is worth noting that cluster #B and C are similar in terms of their other common venues and the only distinguishing factor is the first most common venue

	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Scarborough	3.0	Construction & Landscaping	Bar	Women's Store	Drugstore	Distribution Center	Dive Bar	Dog Run	Doner Restaurant	Donut Shop	Dumpling Restaurant
2	Scarborough	3.0	Mexican Restaurant	Electronics Store	Restaurant	Breakfast Spot	Rental Car Location	Medical Center	Intersection	Bank	Women's Store	Discount Store
3	Scarborough	3.0	Coffee Shop	Mexican Restaurant	Korean BBQ Restaurant	Women's Store	Donut Shop	Discount Store	Distribution Center	Dive Bar	Dog Run	Doner Restaurant
4	Scarborough	3.0	Hakka Restaurant	Athletics & Sports	Bakery	Gas Station	Caribbean Restaurant	Thai Restaurant	Bank	Fried Chicken Joint	Dive Bar	Distribution Center
5	Scarborough	3.0	Playground	Smoke Shop	Jewelry Store	Women's Store	Doner Restaurant	Diner	Discount Store	Distribution Center	Dive Bar	Dog Run
...	...	...	...	...	...	...	...	...	...	...	...	...
92	Etobicoke	3.0	Hardware Store	Tanning Salon	Wings Joint	Kids Store	Fast Food Restaurant	Discount Store	Convenience Store	Gym	Burrito Place	Burger Joint
95	Etobicoke	3.0	Pharmacy	Beer Store	Pizza Place	Coffee Shop	Convenience Store	Café	Shopping Plaza	Liquor Store	Gift Shop	German Restaurant
99	Etobicoke	3.0	Pizza Place	Discount Store	Sandwich Place	Coffee Shop	Chinese Restaurant	Intersection	Dive Bar	Dim Sum Restaurant	Diner	Distribution Center
101	Etobicoke	3.0	Pizza Place	Grocery Store	Pharmacy	Fast Food Restaurant	Sandwich Place	Beer Store	Fried Chicken Joint	Liquor Store	Eastern European Restaurant	Electronics Store
102	Etobicoke	3.0	Garden Center	Rental Car Location	Drugstore	Bar	Women's Store	Dog Run	Diner	Discount Store	Distribution Center	Dive Bar

82 rows × 12 columns

Figure 10: CLUSTER #D

Cluster #D above has no clear label. But what we can see is that it has a lot of variety.

	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
23	North York	4.0	Park	Convenience Store	Women's Store	Donut Shop	Diner	Discount Store	Distribution Center	Dive Bar	Dog Run	Doner Restaurant
25	North York	4.0	Food & Drink Shop	Park	Donut Shop	Diner	Discount Store	Distribution Center	Dive Bar	Dog Run	Doner Restaurant	Women's Store
40	East York	4.0	Park	Convenience Store	Intersection	Women's Store	Donut Shop	Discount Store	Distribution Center	Dive Bar	Dog Run	Doner Restaurant
74	York	4.0	Park	Women's Store	Pool	Donut Shop	Diner	Discount Store	Distribution Center	Dive Bar	Dog Run	Doner Restaurant
98	York	4.0	Park	Women's Store	Donut Shop	Diner	Discount Store	Distribution Center	Dive Bar	Dog Run	Doner Restaurant	Drugstore

Figure 11: CLUSTER #E

Cluster #E above is characterized by parks, women's store and convenience stores.

	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
91	Etobicoke	5.0	Baseball Field	Construction & Landscaping	Women's Store	Distribution Center	Dive Bar	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant
97	North York	5.0	Baseball Field	Women's Store	Distribution Center	Dive Bar	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Diner

Figure 12: CLUSTER #F

Cluster #F has baseball field as its most common venue.

Discussion

To this project, our focus is on cluster #F which has baseball field as its most common theme. As such to answer the question of which neighborhood to open a baseball store, I would prop ose opening up the store in either Etobicoke or North York. Comparing this to our location da ta, this will correspond to:

- Postal code:'M8Y', 'Etobicoke',
  Neighborhood:Old Mill South, King's Mill Park, Sunnylea, Humber Bay, Mimico NE, The Queensway East, Royal York South East, Kingsway Park South East

And,

- Postal code:'M9M', 'North York'
  Neighborhood: Humberlea, Emery

These two locations show the most promise. It is worth noting however that this project made use of location data alone and as such it will be much better to consider other factors pertaining to business requirements to get a more accurate prediction.

Conclusion

Deciding which location to open any kind of store should not be too much of a hassle. Thanks to machine learning, and location data, I was able to determine the most likely place to open a baseball store. Cluster #F shows our result

Bibliography

Toronto. (n.d.). Retrieved 11 23, 2020, from Wikipedia: The Free Encyclopedia: http://en.wikipedia.org/wiki/Toronto