Applied Data Science Capstone

Final Project

Find the best place to open a new brazilian restaurant in Toronto, Canada

Introduction

Assuming that I am a Brazilian immigrant who wants to open a new Brazilian restaurant in Toronto, I need to find out where is the best place to start my enterprise. As I could see when I lived in Brazil, people who like to eat Italian food have a big tendency to like Brazilian food. Therefore, with this assumption, the objective of this project is to define the best place to open my Brazilian restaurant based on the location of existent Italian restaurants.

Business Problem

By using the data analysis and machine learning tools, allied to the clustering machine learning algorithms learned throughout the course I want to answer the question: Where is the best place to consider if I want to open a Brazilian restaurant in Toronto?

Target Audience

Entrepreneurs who wants to find the location to open a reliable Brazilian restaurant in Toronto.

Data

- A list of neighborhoods in Toronto, Canada: the information source is the Wikipedia web site https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M and the data will be extracted from the table present on this site. Example:

Postal Code	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
M8A	Not assigned	Not assigned
M9A	Etobicoke	Islington Avenue, Humber Valley Village
M1B	Scarborough	Malvern, Rouge
M2B	Not assigned	Not assigned
M3B	North York	Don Mills
M4B	East York	Parkview Hill, Woodbine Gardens
M5B	Downtown Toronto	Garden District, Ryerson
M6B	North York	Glencairn
M7B	Not assigned	Not assigned
M8B	Not assigned	Not assigned
M9B	Etobicoke	West Deane Park, Princess Gardens, Martin Grove, Islington, Cloverdale
M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
M2C	Not assigned	Not assigned
M3C	North York	Don Mills
M4C	East York	Woodbine Heights
M5C	Downtown Toronto	St. James Town
M6C	York	Humewood-Cedarvale
M7C	Not assigned	Not assigned
M8C	Not assigned	Not assigned
M9C	Etobicoke	Eringate, Bloordale Gardens, Old Burnhamthorpe, Markland Wood

- The geospatial coordinates of these neighborhoods: the coordinates were obtained from the 'Geospatial_Coordinates.csv' file, provided by the course. Example:

	A	B	C
1	Postal Code	Latitude	Longitude
2	M1B	43.8066863	-79.1943534
3	M1C	43.7845351	-79.1604971
4	M1E	43.7635726	-79.1887115
5	M1G	43.7709921	-79.2169174
6	M1H	43.773136	-79.2394761
7	M1J	43.7447342	-79.2394761
8	M1K	43.7279292	-79.2620294
9	M1L	43.7111117	-79.2845772
10	M1M	43.716316	-79.2394761
11	M1N	43.692657	-79.2648481
12	M1P	43.7574096	-79.273304
13	M1R	43.7500715	-79.2958491
14	M1S	43.7942003	-79.2620294
15	M1T	43.7816375	-79.3043021
16	M1V	43.8152522	-79.2845772
17	M1W	43.7995252	-79.3183887
18	M1X	43.8361247	-79.2056361
19	M2H	43.8037622	-79.3634517
20	M2J	43.7785175	-79.3465557
21	M2K	43.7869473	-79.385975

- Venue data from Foursquare of these neighborhoods: The data was obtained using the Foursquare API to find out where the Italian restaurants are located in Toronto. Example:

	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	649 Variety	43.754513	-79.331942	Convenience Store
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

Methodology

The first step is to to get the list of neighborhoods in Toronto, Canada and remove rows with no borough assigned. Therefore, we need to check if there are inconsistent data like postal codes with assigned to more than one borough and neighborhoods not assigned to any boroughs. The result is our neighborhood data below, with 103 rows, as we can see below:

	Postal Code	Borough	Neighbourhood
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
...	...	...	...
98	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North
99	M4Y	Downtown Toronto	Church and Wellesley
100	M7Y	East Toronto	Business reply mail Processing Centre, South C...
101	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...
102	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...

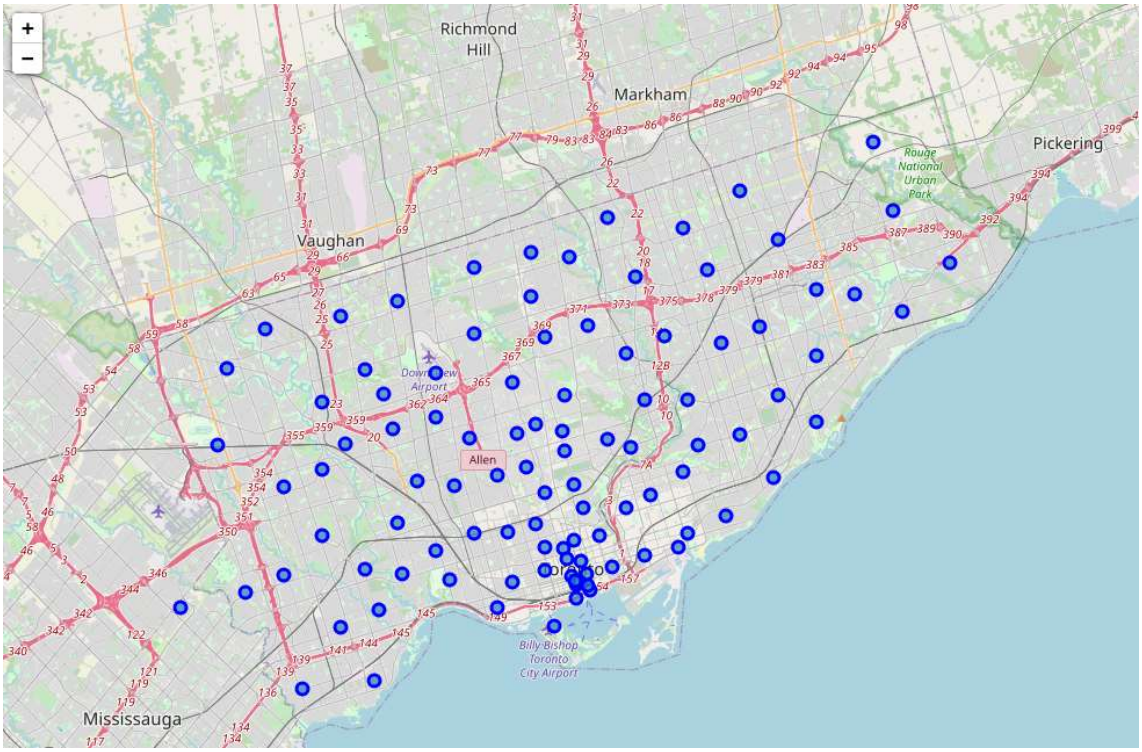
103 rows × 3 columns

Next, we are going to add geospatial coordinates from our data set to the list of neighborhoods:

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
...	...	...	...	...	...
98	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944
99	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160
100	M7Y	East Toronto	Business reply mail Processing Centre, South C...	43.662744	-79.321558
101	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509
102	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999

103 rows × 5 columns

A good way to check if all the geospatial data is correct is to create a map with the points:



Now we can go to Foursquare and get venue data from all the neighborhoods we have, limiting the radius to 500 meters and the total results to 100 per neighborhood. The result set contains 2137 rows of data, as we see below:

	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
...	...	...	...	...	...	...	...
2132	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Koala Tan Tanning Salon & Sunless Spa	43.631370	-79.519006	Tanning Salon
2133	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Once Upon A Child	43.631075	-79.518290	Kids Store
2134	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Value Village	43.631269	-79.518238	Thrift / Vintage Store
2135	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Kingsway Boxing Club	43.627254	-79.526684	Gym

Grouping the result set by category, we have:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Agincourt	5	5	5	5	5	5
Alderwood, Long Branch	9	9	9	9	9	9
Bathurst Manor, Wilson Heights, Downsview North	22	22	22	22	22	22
Bayview Village	4	4	4	4	4	4
Bedford Park, Lawrence Manor East	25	25	25	25	25	25
...	...	...	...	...	...	...
Willowdale, Willowdale East	32	32	32	32	32	32
Willowdale, Willowdale West	6	6	6	6	6	6
Woburn	4	4	4	4	4	4
Woodbine Heights	4	4	4	4	4	4
York Mills West	3	3	3	3	3	3

96 rows × 6 columns

This dataset contains 270 unique categories.

In order to prepare the clustering analysis, we shall analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. “One hot” encoded results:

	Yoga Studio	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...	Trail	Train Station	Vegetarian / Vegan Restaurant	Video Game Store	Vietnamese Restaurant	Ware
0	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	0	
2	0	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	0	
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	

5 rows × 270 columns

Results grouped by the mean of frequency:

	Neighborhood	Yoga Studio	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	...	Trail	Train Station	Vegetarian / Vegan Restaurant	Video Game Store	Vietnamese Restaurant
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	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	0.0	...	0.0	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	0.0	...	0.0	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	0.0	...	0.0	0.0	0.0	0.0	0.00000
4	Bedford Park, Lawrence Manor East	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
91	Willowdale, Willowdale East	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.03125
92	Willowdale, Willowdale West	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000
93	Woburn	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000
94	Woodbine Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000
95	York Mills West	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000

96 rows × 270 columns

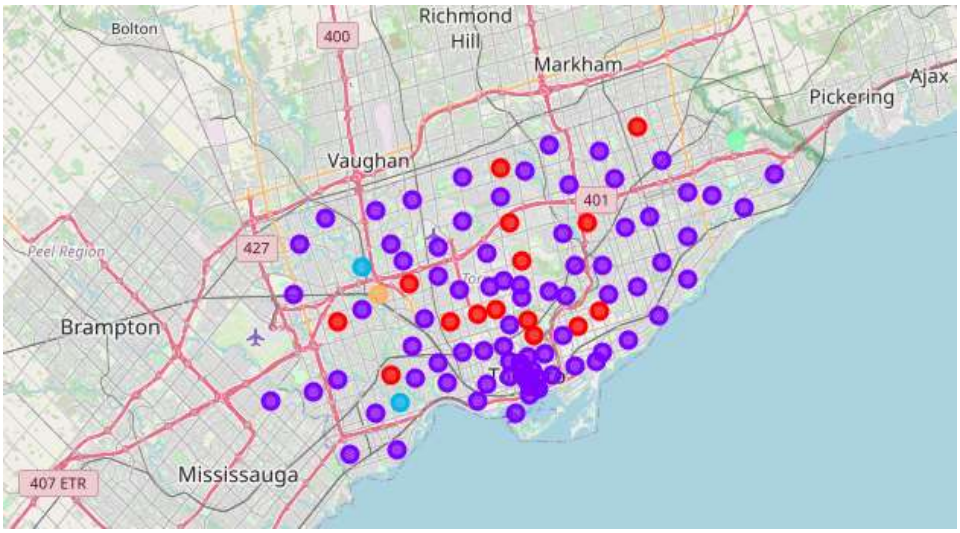
Now, let´s group again the data to find out what are the 10 most common venues for each neighborhood:

	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	Drugstore	Discount Store	Distribution Center	Dog Run	Doner Restaurant
1	Alderwood, Long Branch	Pizza Place	Coffee Shop	Sandwich Place	Athletics & Sports	Pub	Pool	Skating Rink	Gym	Concert Hall	Department Store
2	Bathurst Manor, Wilson Heights, Downsview North	Bank	Coffee Shop	Park	Deli / Bodega	Supermarket	Middle Eastern Restaurant	Sushi Restaurant	Ice Cream Shop	Shopping Mall	Mobile Phone Shop
3	Bayview Village	Café	Bank	Chinese Restaurant	Japanese Restaurant	Women's Store	Diner	Discount Store	Distribution Center	Dog Run	Doner Restaurant
4	Bedford Park, Lawrence Manor East	Italian Restaurant	Coffee Shop	Thai Restaurant	Sandwich Place	Restaurant	Juice Bar	Butcher	Café	Indian Restaurant	Pub

Now we are ready to use the k-means clustering analysis algorithm. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well.

I have clustered the neighborhoods in Toronto into 5 clusters based on their frequency of occurrence for all venues. Based on the results (the concentration of Italian Restaurants in each cluster), I can able to recommend the ideal location to open the Brazilian restaurant.

The image below shows the clusters:



Results

With the clusters defined, we count how many Italian Restaurants we have per cluster. The result is that the cluster number 2 (purple dots) have the most of Italian Restaurants and with this information, we can define the best place to open a Brazilian Restaurant.

Discussion

We said before that the cluster with more Italian Restaurants at the top 10 venues were the cluster number 3. Let´s see the result set of cluster number 2:

	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	North York	1	French Restaurant	Coffee Shop	Hockey Arena	Portuguese Restaurant	Women's Store	Doner Restaurant	Dim Sum Restaurant	Diner	Discount Store	Distribution Center
2	Downtown Toronto	1	Coffee Shop	Park	Café	Pub	Bakery	Breakfast Spot	Theater	Ice Cream Shop	Chocolate Shop	Spa
3	North York	1	Furniture / Home Store	Clothing Store	Vietnamese Restaurant	Coffee Shop	Boutique	Event Space	Accessories Store	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant
4	Downtown Toronto	1	Coffee Shop	College Cafeteria	Sushi Restaurant	Bar	Beer Bar	Smoothie Shop	Sandwich Place	Café	Portuguese Restaurant	Persian Restaurant
7	North York	1	Gym	Coffee Shop	Clothing Store	Japanese Restaurant	Beer Store	Restaurant	Chinese Restaurant	Supermarket	Discount Store	Café
...	...	...	...	...	...	...	...	...	...	...	...	...
96	Downtown Toronto	1	Coffee Shop	Café	Bakery	Chinese Restaurant	Italian Restaurant	Pizza Place	Restaurant	Pub	General Entertainment	Snack Place
97	Downtown Toronto	1	Coffee Shop	Café	Hotel	Restaurant	Gym	Japanese Restaurant	American Restaurant	Seafood Restaurant	Steakhouse	Asian Restaurant
99	Downtown Toronto	1	Coffee Shop	Japanese Restaurant	Sushi Restaurant	Restaurant	Gay Bar	Café	Bubble Tea Shop	Pub	Yoga Studio	Mediterranean Restaurant
100	East Toronto	1	Yoga Studio	Garden Center	Skate Park	Light Rail Station	Farmers Market	Fast Food Restaurant	Spa	Burrito Place	Butcher	Restaurant
102	Etobicoke	1	Grocery Store	Kids Store	Bakery	Discount Store	Tanning Salon	Burrito Place	Burger Joint	Sandwich Place	Thrift / Vintage Store	Supplement Shop

81 rows × 12 columns

As we can see, the cluster number 2 has too many different boroughs, so let´s see in which borough Italian Restaurant appears as the 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
55	North York	1	Italian Restaurant	Coffee Shop	Thai Restaurant	Sandwich Place	Restaurant	Juice Bar	Butcher	Café	Indian Restaurant	Pub

Based on this information, we can define that the North York Borough is a good place to open a Brazilian Restaurant

Conclusion

With the assumption that people who like to eat Italian Food and after the analysis of the clustering data, obtained with a dataset created with Foursquare venue data, we came to the conclusion that the best borough in Toronto, Canada, to open a Brazilian Restaurant is North York.