Capstone Project - The Battle of the Neighborhoods

1) Background

One of the most common businesses is that of a restaurant. There are many reasons why this is the case, chief among these is that it is an easy business to get into. The qualifications for opening a restaurant are low. If an individual could scrape together the necessary funds, he/she could open a new restaurant. This is one reason why so many immigrants open restaurants. For this project, I am assuming that a client is interested in finding a good Toronto neighborhood to open a Japanese restaurant in.

2) Business Problem

While it is not difficult to set up a restaurant, it is not easy to successfully run one. As any real estate agent will be quick to say, the three most important variables in having a successful venture is location, location, location. The client can serve fantastic Japanese food and provide unparalleled service, but the restaurant may still fail if it is in a poor location. For instance, the client may face an uphill battle if the restaurant in a neighborhood with many similar restaurants. These entrenched competitors will do all that they could to take customers from the client. Armed with the knowledge that a good location is vital to the success of the client's venture; I will analyze Toronto's geographic data to find an ideal neighborhood for the hypothetical client's new Japanese restaurant.

3) Data

For this project, I will aggregate and analyze data from three sources: Toronto neighborhood data from Wikipedia, latitude and longitude data from Geocoder, and venue data from Foursquare.

Firstly, I imported postal code, borough, and neighborhood data from Wikipedia. I then merge the resulting dataframe with latitude and longitude data from Geocoder. I call the resulting dataframe toronto. Here are the first few observations of dataframe toronto:

	Postalcode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	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
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	East Birchmount Park, Ionview, Kennedy Park	43.727929	-79.262029
7	M1L	Scarborough	Clairlea, Golden Mile, Oakridge	43.711112	-79.284577
8	M1M	Scarborough	Cliffcrest, Cliffside, Scarborough Village West	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff, Cliffside West	43.692657	-79.264848

I next acquire venue data from Foursquare. Here are the first few observations of the venue data acquire from Foursquare:

	name	categories	lat	lng
0	Toronto Pan Am Sports Centre	Athletics & Sports	43.790623	-79.193869
1	African Rainforest Pavilion	Zoo Exhibit	43.817725	-79.183433
2	Toronto Zoo	Zoo	43.820582	-79.181551
3	Polar Bear Exhibit	Zoo	43.823372	-79.185145
4	Australasia Pavillion	Zoo Exhibit	43.822563	-79.183286

The last two data processing steps are to merge the venue data with the geographic data and to one hot encode venue data for the k-means clustering algorithm to be able to complete the task.

4) Methodology

After merging geographic data with venue data, I create a new dataframe that lists the ten most common types of venues for each neighborhood. Here are the first five observations in the dataframe:

	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	Adelaide, King, Richmond	Coffee Shop	Café	Steakhouse	Bar	American Restaurant	Bakery	Sushi Restaurant	Asian Restaurant	Thai Restaurant	Restaurant
1	Agincourt	Lounge	Breakfast Spot	Latin American Restaurant	Skating Rink	Women's Store	Eastern European Restaurant	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant
2	Agincourt North, L'Amoreaux East, Milliken, St...	Playground	Park	Dumpling Restaurant	Diner	Discount Store	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Eastern European Restaurant
3	Albion Gardens, Beaumond Heights, Humbergate, ...	Fast Food Restaurant	Fried Chicken Joint	Sandwich Place	Discount Store	Pizza Place	Beer Store	Japanese Restaurant	Pharmacy	Grocery Store	Gift Shop
4	Alderwood, Long Branch	Pizza Place	Gym	Pool	Sandwich Place	Athletics & Sports	Pub	Skating Rink	Coffee Shop	Pharmacy	Dumpling Restaurant

Restaurants are some of the most common venues in all five of these neighborhoods. This suggests that restaurants do well in Toronto overall. Particularly interesting for the client is the fact that Asian restaurants are popular in many of Toronto's neighborhoods. For example, the 7^{th}, 8^{th}, and 9^{th} most common venues in the Adelaide, King, Richmond neighborhood are Asian restaurants. This could be a double-edged sword, however. An abundance of Asian restaurants indicate that Asian food is in high demand. But it also means that there will be many entrenched competitors for the client. To solve this conundrum, I will use the k-means clustering algorithm to identify neighborhood clusters where Asian food is in particularly high demand. To avoid the market oversaturation problem, I will only consider neighborhoods within these clusters where there are few Asian restaurants.

5) Results

The k-means algorithm is an unsupervised algorithm, meaning that the modeler cannot define the criteria for the algorithm. The modeler can simply interpret the clusters that the algorithm generates. Some of the clusters that the algorithm returns are not particularly useful. For example, the algorithm returned several clusters that have only one neighborhood. These unique neighborhoods simply do not have comparable neighborhoods. On the other side of the spectrum, the algorithm returned a cluster that included nearly every neighborhood in the dataframe. It is difficult to derive the characteristics of such a cluster. Some of the clusters are easier to interpret. I will present these clusters.

Here is an example of a cluster that the algorithm generated:

	Neighborhood	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	York Mills West	2.0	Park	Bank	Convenience Store	Women's Store	Eastern European Restaurant	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant
25	Parkwoods	2.0	Park	Food & Drink Shop	Women's Store	Dumpling Restaurant	Discount Store	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Electronics Store
40	East Toronto	2.0	Park	Metro Station	Convenience Store	Women's Store	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Electronics Store
74	Caledonia-Fairbanks	2.0	Park	Women's Store	Market	Fast Food Restaurant	Grocery Store	Dessert Shop	Event Space	Ethiopian Restaurant	Empanada Restaurant	Electronics Store
90	The Kingsway, Montgomery Road, Old Mill North	2.0	River	Park	Women's Store	Drugstore	Diner	Discount Store	Dog Run	Doner Restaurant	Donut Shop	Dumpling Restaurant
98	Weston	2.0	Park	Women's Store	Eastern European Restaurant	Discount Store	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Electronics Store

The algorithm clustered neighborhoods that are dominated by recreational facilities such as parks, rivers, and dog runs, and retailers, such as women's stores. Restaurants are popular in this cluster as well. However, this cluster does not satisfy the first criteria, which is being home to numerous Asian restaurants.

	Neighborhood	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
6	East Birchmount Park, Ionview, Kennedy Park	8.0	Discount Store	Department Store	Bus Station	Coffee Shop	Dumpling Restaurant	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Women's Store
13	Clarks Corners, Sullivan, Tam O'Shanter	8.0	Pharmacy	Pizza Place	Shopping Mall	Bank	Italian Restaurant	Thai Restaurant	Chinese Restaurant	Noodle House	Fried Chicken Joint	Fast Food Restaurant
15	L'Amoreaux West	8.0	Fast Food Restaurant	Chinese Restaurant	Pharmacy	Grocery Store	Pizza Place	Sandwich Place	Breakfast Spot	Coffee Shop	Bubble Tea Shop	Thrift / Vintage Store
24	Willowdale West	8.0	Pizza Place	Pharmacy	Discount Store	Coffee Shop	Drugstore	Diner	Dog Run	Doner Restaurant	Donut Shop	Eastern European Restaurant
89	Alderwood, Long Branch	8.0	Pizza Place	Gym	Pool	Sandwich Place	Athletics & Sports	Pub	Skating Rink	Coffee Shop	Pharmacy	Dumpling Restaurant
99	Westmount	8.0	Pizza Place	Middle Eastern Restaurant	Sandwich Place	Discount Store	Chinese Restaurant	Coffee Shop	Intersection	Diner	Dog Run	Doner Restaurant
101	Albion Gardens, Beaumond Heights, Humbergate, ...	8.0	Fast Food Restaurant	Fried Chicken Joint	Sandwich Place	Discount Store	Pizza Place	Beer Store	Japanese Restaurant	Pharmacy	Grocery Store	Gift Shop

The cluster with the most potential is listed above. Restaurants dominate this cluster of neighborhoods. Asian cuisine is in the top 10 in most neighborhoods, fulfilling the first criteria and suggesting that a new Asian restaurant may do well in this cluster of neighborhoods. Within the cluster of neighborhoods where Asian food is in demand, the second criteria is to identify a neighborhood with few Asian food options. The Willowdale neighborhood is the only neighborhood in this cluster without an Asian restaurant in the top 10. This neighborhood is a good prospect for the client to build a new Japanese restaurant. Further reviewing Willowdale West shows that the neighborhood only has three venues, 1 coffee shop, one discount store, and 1 pizza place. There is a dearth of sitdown restaurants. Willowdale West is in a neighborhood cluster where Asian food is popular but has no Asian restaurants itself. This neighborhood is where the client should build a new Japanese restaurant.

6) Discussion

The k-means algorithm is a very useful algorithm for solving problems such as identifying a good neighborhood to place a new restaurant. However, it does have limitations. Because it is an unsupervised algorithm, the user could not define criteria. This resulted in many clusters that were not useful. For example, there were several clusters that had only one neighborhood. It is difficult to comprehend the characteristic of the cluster when the cluster only has one neighborhood. This problem required me to run the algorithm several times before finding a neighborhood that satisfied both criteria. There are likely other algorithms that could generate similar if not better results without the limitation discribed. Finding an algorithm better suited for this analysis could be an area of further research.

The recommendation that I make for the client is to build the new Japanese restaurant in the Willowdale West neighborhood. The neighborhood is in a neighborhood cluster where Asian food is in demand and in a neighborhood with zero Asian restaurants.

7) Conclusion

In this project, I searched for a neighborhood for a fictional client to build a new Japanese restaurant in. While it is not difficult to start a new restaurant, it is incredibly difficult to run a successful one. To improve the chances of the restaurant's success, the client would need to find a place where Japanese food is in demand but competition is limited, or non-existent.

To find a good neighborhood to build a new Japanese restaurant in, I utilized Toronto neighborhood data from Wikipedia, longitude and latitude data from Geocoder and venue data from Foursquare. I apply the k-means algorithm to identify a neighborhood that satisfied two criteria:

1) In a neighborhood cluster with many Asian restaurants, indicating that Asian food is in demand.
2) In a neighborhood that has few or zero Asian restaurants, indicating that competition is limited.

The Willowdale West neighborhood satisfied both criteria. It is in a neighborhood cluster where there are Asian restaurants and host no Asian restaurant of its own; and therefore, is an ideal neighborhood for the client to build a new Japanese restaurant.

Lastly, an area for additional exploration is to find one or more additional algorithms that could answer the same question. The k-means algorithm was time consuming because it was unsupervised. Furthermore, having additional algorithms could validate or disprove the conclusion that Willowdale West is a good candidate for a new Japanese restaurant.