Coursera Capstone Project

Zhonghao Sun

Introduction

Toronto is the capital of Canada. An international supermarket chain company is planning to open its first supermarket in Toronto. The aim of this project is to find a proper neighborhood location for its first store. A number of factors need to be considered to determine a proper location, these include demand, competition, and reputation for future development. The company may have different development strategies. For example, a less competitive area may be not good for the company to improve its reputation for future development, while it may be difficult to be profit in a more competitive area. The task for a data scientist is to collect and analyze Toronto neighborhoods' data and provide statistical information for the board to make decision. Specifically, the following need to be explored:

1) Compare population (demand), number of supermarket and grocery stores (competition), number of other venues such as restaurants, bars, hotels, schools that can increase demand,
2) Cluster neighborhoods and find the current development patterns,
3) Identify proper neighborhoods based on different development strategies of the company.

The target audience is the marketing and development department of the supermarket company.

Data

To solve the problem, we need to know neighborhood name, geographical information, population, and venue data.

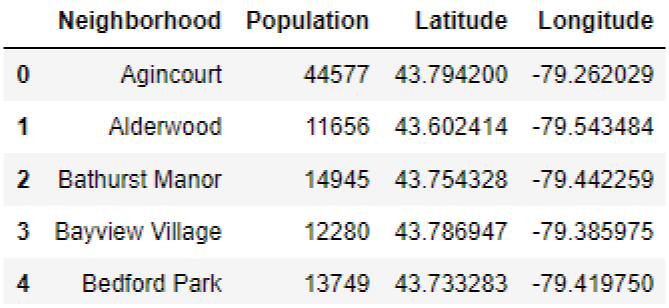
1) Neighborhood name and population. The name and population of Toronto neighborhoods can be find in Wikipedia: https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods. It contains a table list. We can use web scraping techniques to extract the data with pandas and beautifulsoup packages. The data is stored in a dataframe and can be further processed.
2) Geographical information of neighborhoods. The geographical information is required to request venue data. The latitude and longitude coordinates of neighborhoods can be

obtained using Geocoder package or the csv file from the class. This will be a table list of neighborhoods name, and their latitude and longitude coordinates.

3) Venue data. We use Foursquare API to get venue data for neighborhoods. Foursquare.com is one of the largest databases of venues. We can obtain a table list of neighborhoods and nearby venues, including venue name and category by using the Foursquare API.

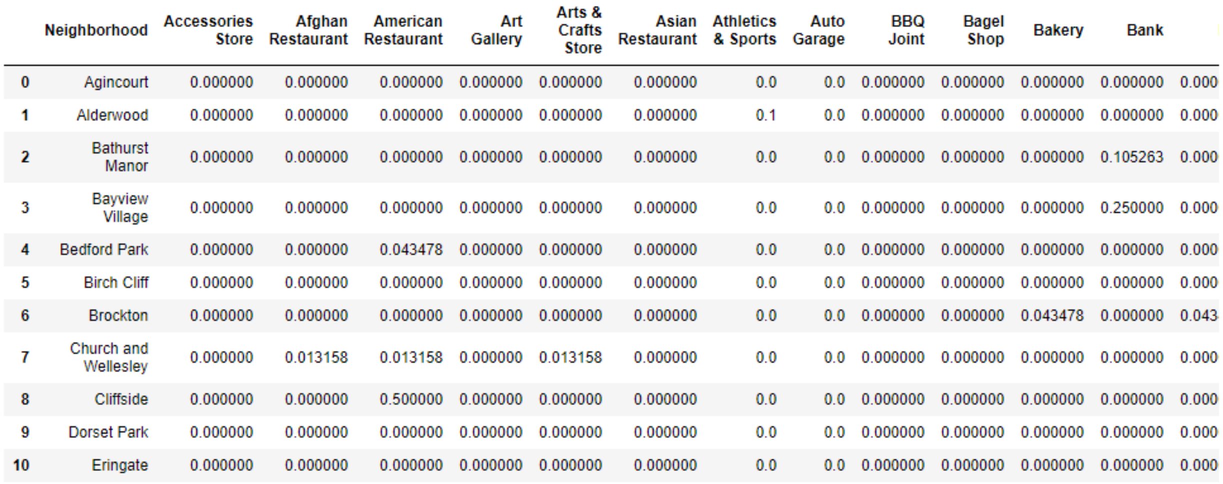
Methodology

First, we obtain neighborhood names, population, and geographical information.



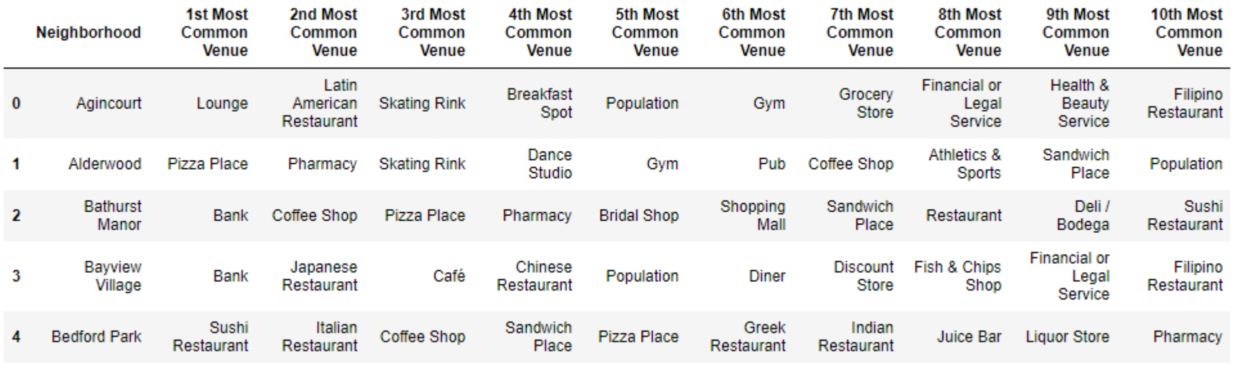
	Neighborhood	Population	Latitude	Longitude
0	Agincourt	44577	43.794200	-79.262029
1	Alderwood	11656	43.602414	-79.543484
2	Bathurst Manor	14945	43.754328	-79.442259
3	Bayview Village	12280	43.786947	-79.385975
4	Bedford Park	13749	43.733283	-79.419750

Then we get nearby venues using the Foursquare API.



	Neighborhood	Accessories Store	Afghan Restaurant	American Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Garage	BBQ Joint	Bagel Shop	Bakery	Bank	
0	Agincourt	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000
1	Alderwood	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.1	0.0	0.000000	0.000000	0.000000	0.000000	0.000
2	Bathurst Manor	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.105263	0.000
3	Bayview Village	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.250000	0.000
4	Bedford Park	0.000000	0.000000	0.043478	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000
5	Birch Cliff	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000
6	Brockton	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.043478	0.000000	0.043
7	Church and Wellesley	0.000000	0.013158	0.013158	0.000000	0.013158	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000
8	Cliffside	0.000000	0.000000	0.500000	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000
9	Dorset Park	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000
10	Eringate	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000

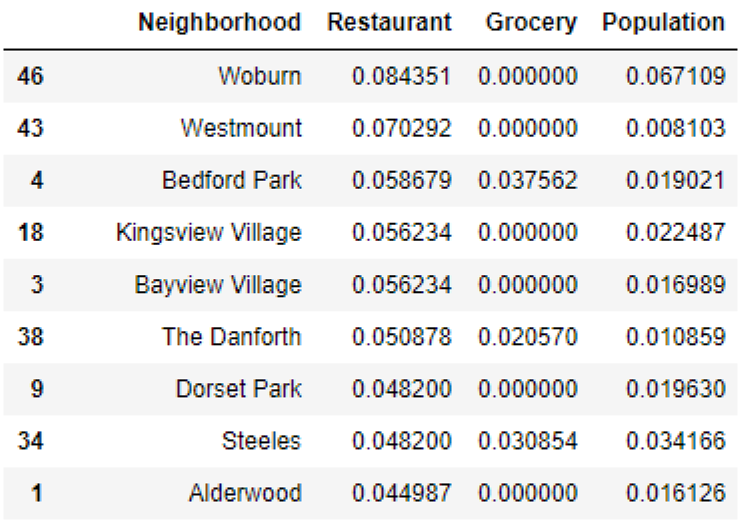
We can find the 10 most common venues in each neighborhood.



	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	Latin American Restaurant	Skating Rink	Breakfast Spot	Population	Gym	Grocery Store	Financial or Legal Service	Health & Beauty Service	Filipino Restaurant
1	Alderwood	Pizza Place	Pharmacy	Skating Rink	Dance Studio	Gym	Pub	Coffee Shop	Athletics & Sports	Sandwich Place	Population
2	Bathurst Manor	Bank	Coffee Shop	Pizza Place	Pharmacy	Bridal Shop	Shopping Mall	Sandwich Place	Restaurant	Deli / Bodega	Sushi Restaurant
3	Bayview Village	Bank	Japanese Restaurant	Café	Chinese Restaurant	Population	Diner	Discount Store	Fish & Chips Shop	Financial or Legal Service	Filipino Restaurant
4	Bedford Park	Sushi Restaurant	Italian Restaurant	Coffee Shop	Sandwich Place	Pizza Place	Greek Restaurant	Indian Restaurant	Juice Bar	Liquor Store	Pharmacy

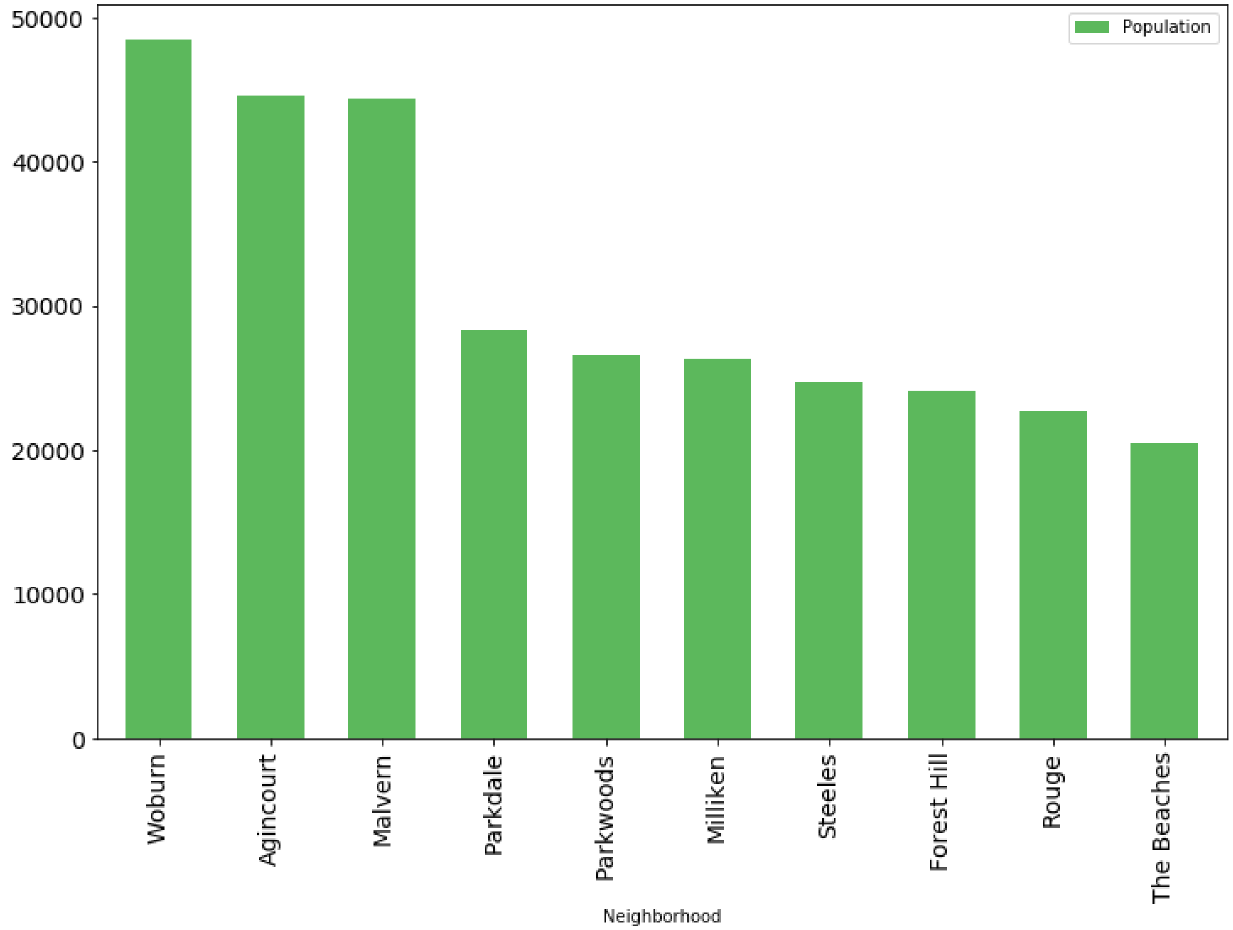
Note that population is considered because a large population may mean a high demand.

To address the demand and competition factors, we consider that restaurants can attract people and can increase the profit of the supermarket, while grocery stores, other supermarkets, and shopping malls will increase the competition. We collect these data from the venue data and normalize:

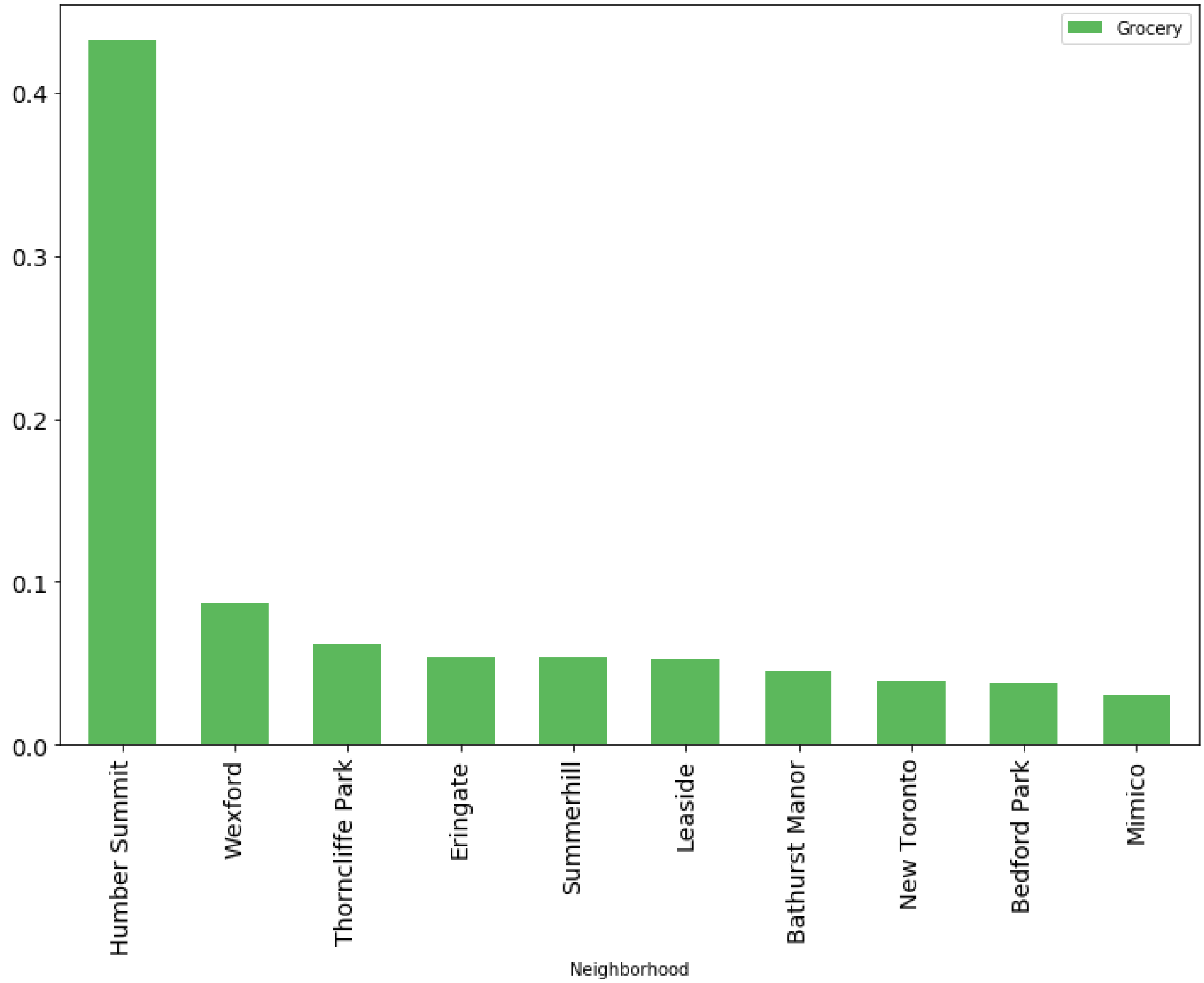


	Neighborhood	Restaurant	Grocery	Population
46	Woburn	0.084351	0.000000	0.067109
43	Westmount	0.070292	0.000000	0.008103
4	Bedford Park	0.058679	0.037562	0.019021
18	Kingsview Village	0.056234	0.000000	0.022487
3	Bayview Village	0.056234	0.000000	0.016989
38	The Danforth	0.050878	0.020570	0.010859
9	Dorset Park	0.048200	0.000000	0.019630
34	Steeles	0.048200	0.030854	0.034166
1	Alderwood	0.044987	0.000000	0.016126

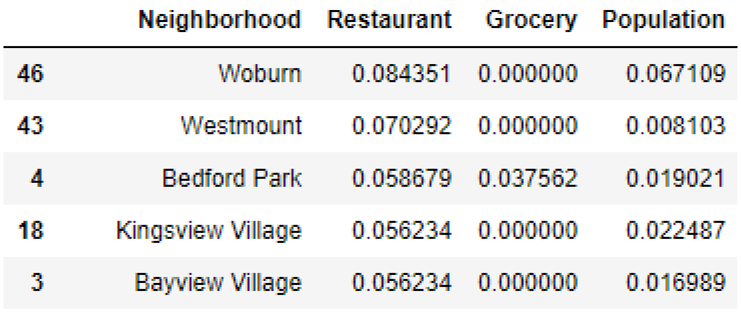
We can find the neighborhoods with highest population:



We can also find the neighborhood with the largest number of grocery stores, supermarket, and shopping malls:

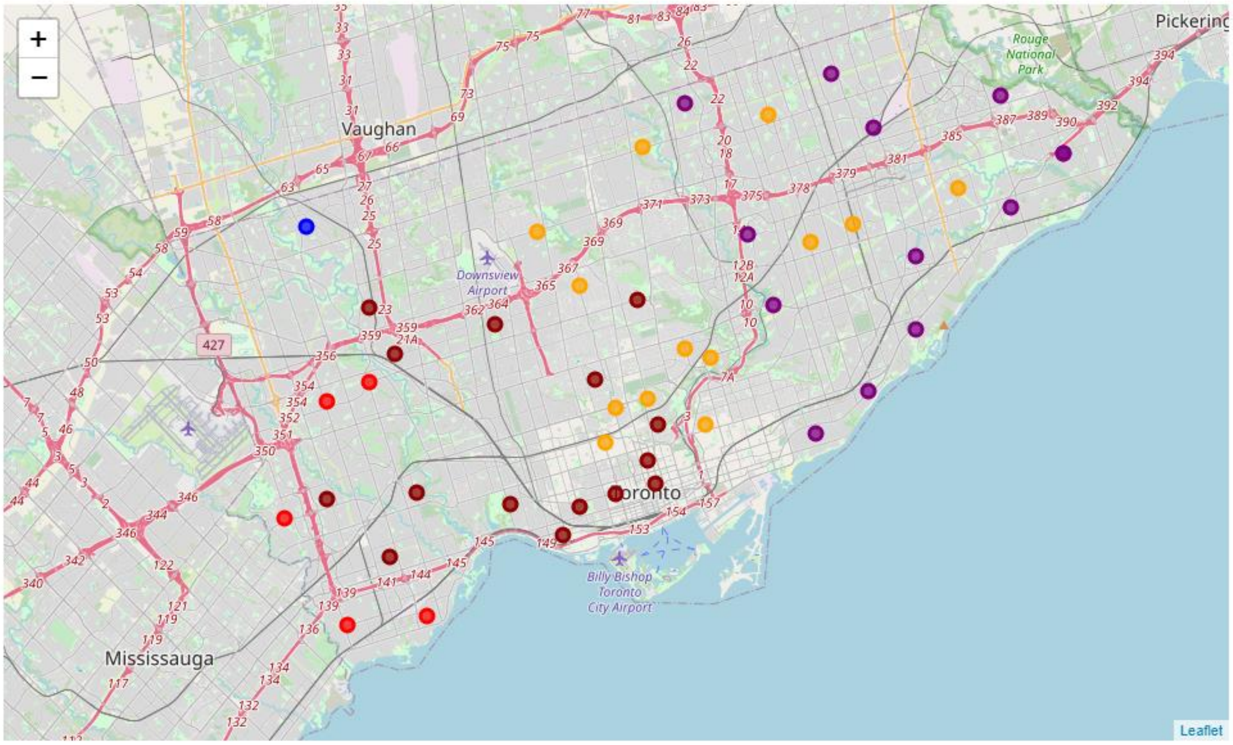


And neighborhoods with many restaurants:



	Neighborhood	Restaurant	Grocery	Population
46	Woburn	0.084351	0.000000	0.067109
43	Westmount	0.070292	0.000000	0.008103
4	Bedford Park	0.058679	0.037562	0.019021
18	Kingsview Village	0.056234	0.000000	0.022487
3	Bayview Village	0.056234	0.000000	0.016989

Next, we perform k-means clustering. We choose 3 clusters



Results

The five neighborhoods which have the highest population are Woburn, Agincourt, Malvern, Parkdale, and Parkwoods. The five neighborhoods which have the largest number of restaurants are Woburn, Westmount, Bedford Park, Kingsview Village, and Bayview Village. The five neighborhoods which have the largest number of grocery stores, supermarkets, and shopping malls are Humber Summit, Wexford, Thorncliffe Park, Eringate, and Summerhill.

We identify 5 clusters:

(1) Cluster #1 (darkred) has a relative low score of restaurant and grocery, and a relative low score of population.

(2) Cluster #2 (red) has a relative high score of restaurants, and a relative low score of population.

(3) Cluster #3 (blue) has a very high score of restaurants, and a very low score of population. There is only one neighborhood in this cluster: Humber Summit.

(4) Cluster #4 (purple) has a very low score of restaurant/grocery, and a relative high score of population, such as Bayview Village, Bedford Park, Dorset Park, and Woburn.

(5) Cluster #5 (orange) has a relative high score of restaurants.

Discussion

We can propose several neighborhoods based on the different development strategies of the company. Cluster #1 has low population and is not considered.

Strategies	Neighborhoods
Highest competition	Cluster #3: Humber Summit
Around restaurants, high competition	Cluster #2: Eringate, New Toronto Cluster #5: Bathurst Manor, Bedford Park
Around restaurants, low competition	Cluster #2: Alderwood, Kingsview Village Cluster #5: Bayview Village, Dorset Park, Woburn
High population, low competition	Cluster #4: Agincourt, Malvern, Parkwoods

Conclusion

In conclusion, we analyzed the population and nearby venue data of Toronto neighborhoods. We find neighborhoods with the highest population, the largest number of supermarkets. We identify 5 clusters of neighborhoods based on the level of demand and competition. Finally, we recommend a few neighborhoods for the new supermarket that can satisfy different development strategies of the company.