Capstone Project: The Battle of Neighborhoods

Opening a Sporting Goods Shop in Toronto

By: M Farshchin

Introduction/Business Problem:

Toronto is one of the largest metropolitan cities in Canada. As the most populous city in Canada, this city is an attractive destination for many businesses. Toronto consists of 10 Boroughs and 103 neighborhoods. This project aims at exploring Toronto neighborhoods and finding potential locations for opening a **sporting goods shop** in a shopping mall or shopping plaza. For this purpose, a data-driven approach is applied to make an informed decision. As discussed in the following sections, different sources are used to obtain data about each neighborhood and K-Mean clustering method is used to cluster similar neighborhoods. The results of this study could be used by distributors of sports goods that are interested in opening new shops in the Toronto.

Data Description:

For this problem the following data sources are used:

- Data about Toronto neighborhoods are scraped from the following Wikipedia page: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

This dataset provides information about different postal areas of Toronto along with their related Borough and neighborhoods.

- Geographical coordinates of each postal code is obtained from the following csv file: https://cocl.us/Geospatial_data

  This dataset provides geographical coordinates (Latitude and Longitude) of different postal areas of Toronto.
- Foursquare API is used to get all venues for each neighborhoods of Toronto.

Methodology:

In order to find the potential locations for opening a new **sport goods shop** in Toronto, the following criterion are considered:

1- Neighborhoods with higher density of sport venues are more attractive to open the new shop because there is more demand in such neighborhoods.

2- Neighborhoods that already have at least one sport goods shop are not suitable because there will be more competition.

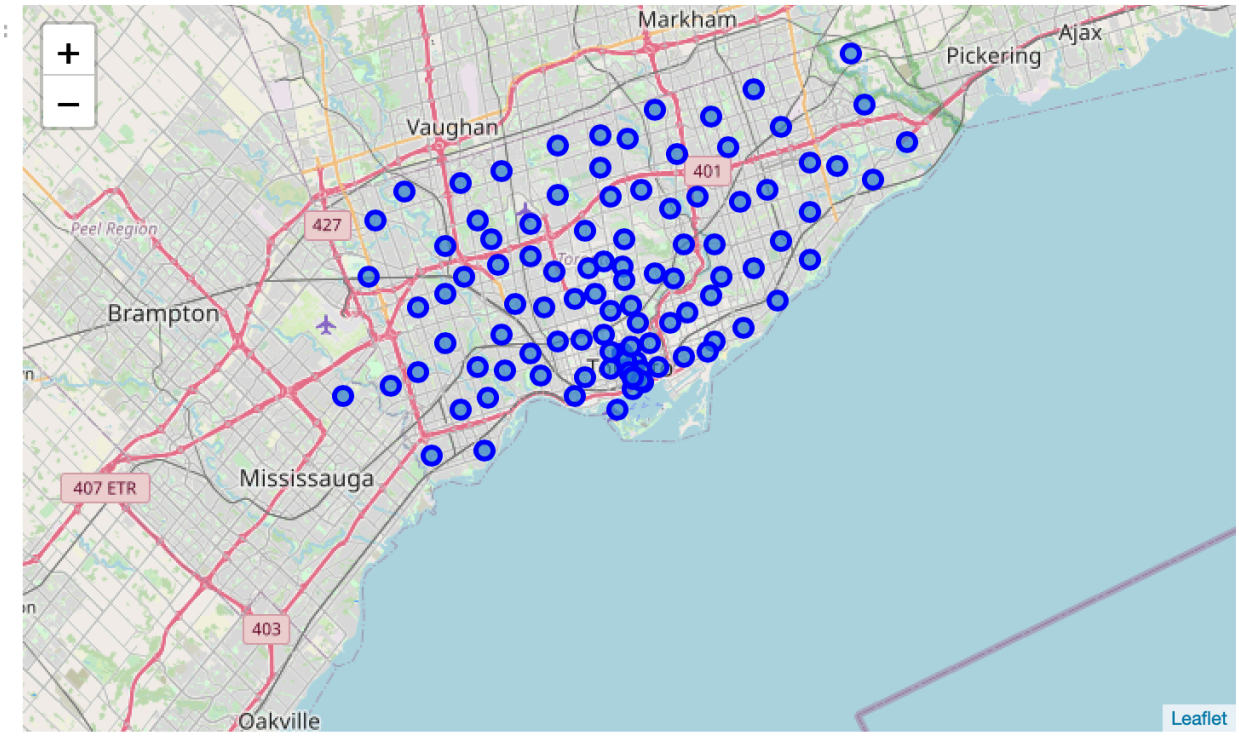
3- In this project we assume that the customer is interested in opening the sport goods shop in a shopping mall or shopping plaza. Therefore, the neighborhoods with more shopping malls or shopping plazas are more attractive because they give more options.

Preprocessing the data and exploratory data analysis:

First step of the project was to collect and organize the required data. As mentioned in the Data Description section, information about different neighborhoods of Toronto was collected and processed into a data frame. A snapshot of the first few rows of the resulting data frame is shown in the following image. This data frame includes information such ‘Postalcode’, ‘Borough’, ‘Latitude’, and ‘Longitude’ of each neighborhood of Toronto. Exploring this data frame shows that Toronto has 10 Boroughs and 103 Neighborhoods.

	Postalcode	Borough	Neighborhood	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

Having this information, the following map of Toronto with neighborhoods superimposed on the top is created.



In the next step of this project, the **Foursquare API** is used to collect information about the different venues of different neighborhoods of Toronto. The following figure shows the main categories of Foursquare data base along with their category id:

```
4d4b7104d754a06370d81259 Arts & Entertainment
4d4b7105d754a06372d81259 College & University
4d4b7105d754a06373d81259 Event
4d4b7105d754a06374d81259 Food
4d4b7105d754a06376d81259 Nightlife Spot
4d4b7105d754a06377d81259 Outdoors & Recreation
4d4b7105d754a06375d81259 Professional & Other Places
4e67e38e036454776db1fb3a Residence
4d4b7105d754a06378d81259 Shop & Service
4d4b7105d754a06379d81259 Travel & Transport
```

Since we are interested in finding potential locations for opening a Sporting Goods Shop in a shopping mall or shopping plaza, we need to obtain information about venues in two main categories of **Outdoors & Recreation** and **Shop & Service**.

The following figure shows some of the subcategories of the Outdoors & Recreation:

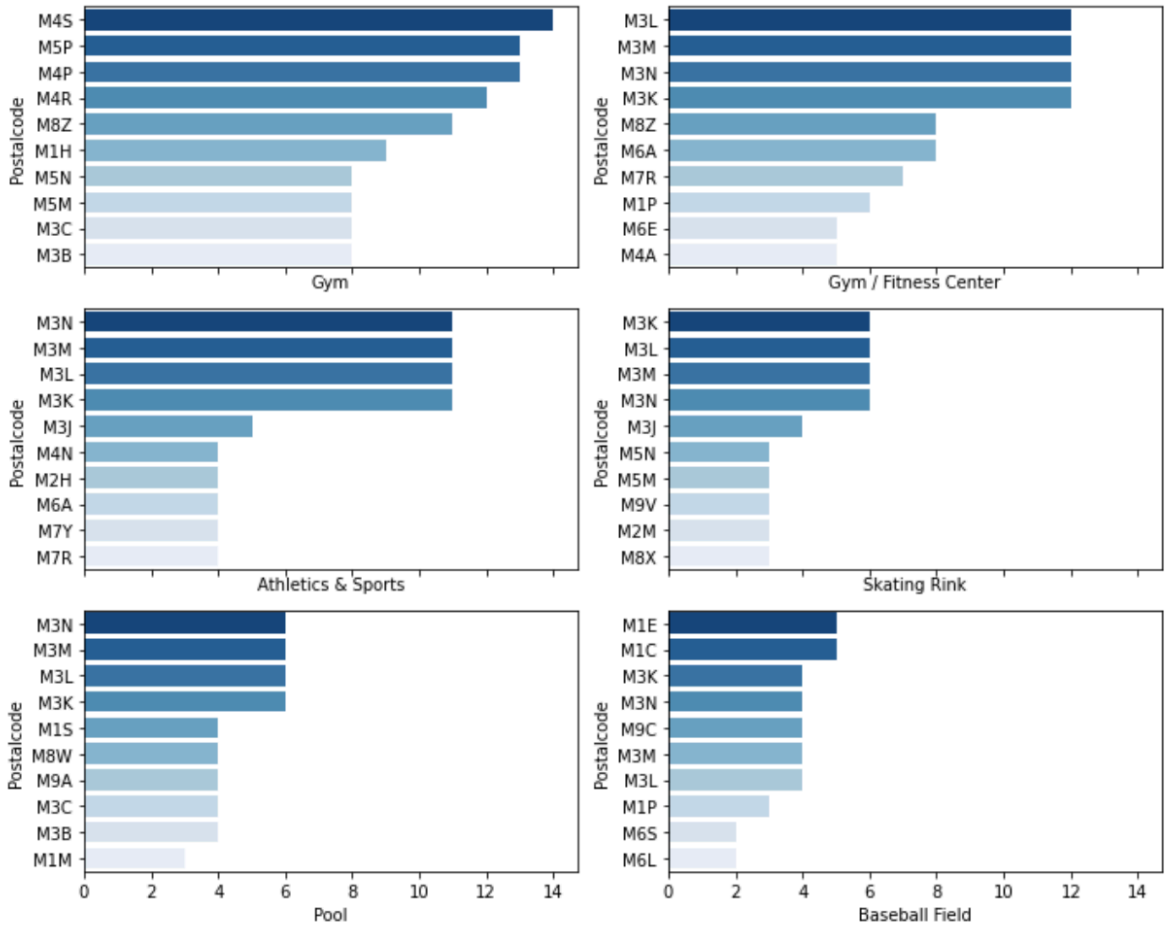
```
[15]: {'4d4b7105d754a06377d81259': 'Outdoors & Recreation',
       '4f4528bc4b90abdf24c9de85': 'Athletics & Sports',
       '52e81612bcbc57f1066b7a2b': 'Badminton Court',
       '4bf58dd8d48988d1e8941735': 'Baseball Field',
       '4bf58dd8d48988d1e1941735': 'Basketball Court',
       '52e81612bcbc57f1066b7a2f': 'Bowling Green',
       '56aa371be4b08b9a8d57351a': 'Curling Ice',
       '4bf58dd8d48988d1e6941735': 'Golf Course',
       '58daa1558bbb0b01f18ec1b0': 'Golf Driving Range',
       '4bf58dd8d48988d175941735': 'Gym / Fitness Center',
       '52f2ab2ebcbc57f1066b8b47': 'Boxing Gym',
       '503289d391d4c4b30a586d6a': 'Climbing Gym',
       '52f2ab2ebcbc57f1066b8b49': 'Cycle Studio',
```

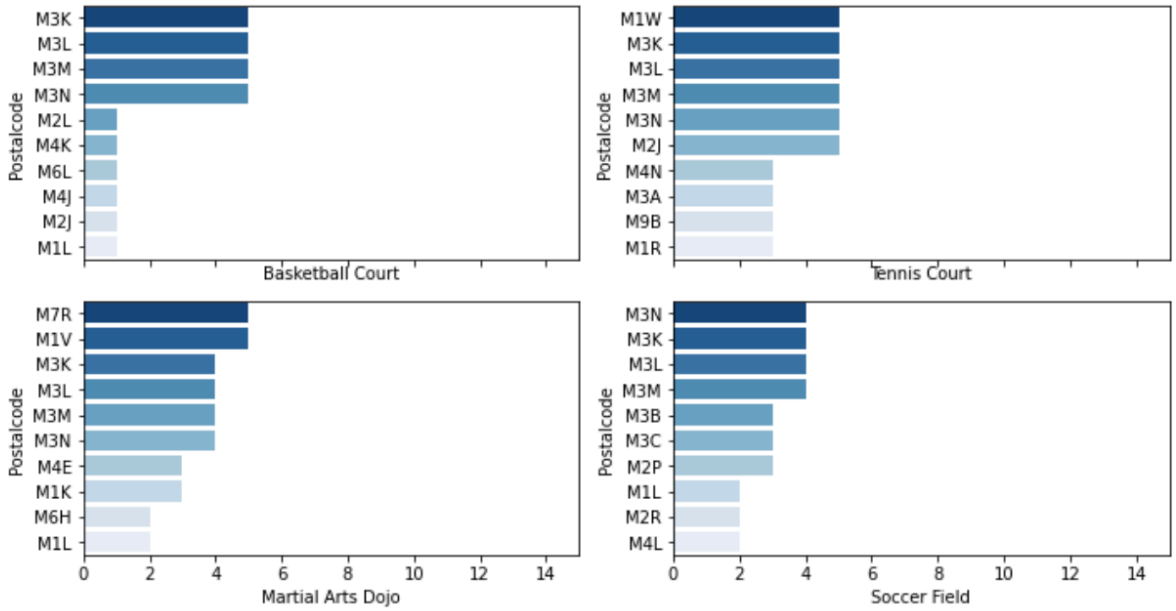
From this category, we get all venues that are in a radius of about one mile (~1610 m) from the center of each neighborhood. In addition, since we are interested in only sport venues in this category, we only keep the sport venues. A snapshot of the resulting data frame is shown below.

	index	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	1	Parkwoods	43.753259	-79.329656	GoodLife Fitness North York Victoria Terrace	43.742128	-79.314590	Gym
1	2	Parkwoods	43.753259	-79.329656	GH20	43.765608	-79.342918	Gym / Fitness Center
2	4	Parkwoods	43.753259	-79.329656	Pheasant Run Golf Course	43.758386	-79.337191	Golf Course
3	5	Parkwoods	43.753259	-79.329656	West Ellesmere Community Gym	43.758155	-79.307618	Athletics & Sports
4	8	Parkwoods	43.753259	-79.329656	Parkway Valley Tennis Club	43.754481	-79.318285	Tennis Court

To get a better understanding of the obtained venues, the top 10 venues are extracted and analyzed. The results are shown below.

	count	mean	std	min	25%	50%	75%	max
Gym	98.0	4.265306	2.797462	0.0	2.0	4.0	5.0	14.0
Gym / Fitness Center	98.0	2.510204	2.057147	0.0	1.0	2.0	4.0	12.0
Athletics & Sports	98.0	1.418367	1.630404	0.0	0.0	1.0	2.0	11.0
Skating Rink	98.0	1.030612	0.989159	0.0	0.0	1.0	1.0	6.0
Pool	98.0	1.244898	1.175885	0.0	0.0	1.0	2.0	6.0
Baseball Field	98.0	0.795918	1.093176	0.0	0.0	0.0	1.0	5.0
Basketball Court	98.0	0.224490	0.618122	0.0	0.0	0.0	0.0	5.0
Tennis Court	98.0	0.806122	1.145661	0.0	0.0	0.0	1.0	5.0
Martial Arts Dojo	98.0	0.826531	1.055428	0.0	0.0	1.0	1.0	5.0
Soccer Field	98.0	0.428571	0.799484	0.0	0.0	0.0	1.0	4.0





The resulting figures show that Gyms and Gym/Fitness centers are the most popular venues in Toronto.

Regarding the shopping venues, since our assumption is that the customer is looking for available shopping malls and shopping plazas to open the new shop, in the next step, we get the information about availability of these venues in neighborhoods of Toronto. In addition, our goal is to avoid neighborhoods that already have a 'Sport Goods Shop', therefore we get venues for this category as well.

Now we summarize all information into the following data frame (only first few rows are shown).

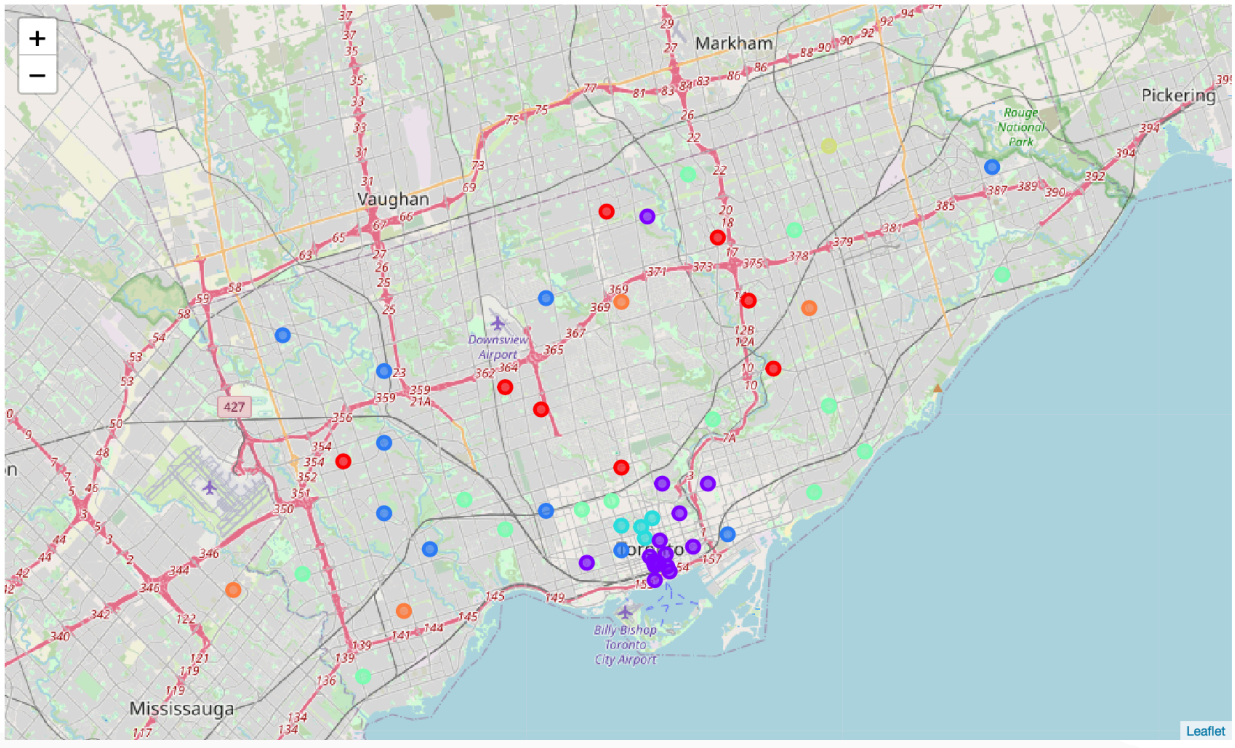
	Neighborhood	# of sport venues	# of shopping venues	# of Sporting Goods Shop
0	Agincourt	22	5	1
1	Alderwood, Long Branch	17	3	0
2	Bathurst Manor, Wilson Heights, Downsview North	13	1	0
3	Bayview Village	8	4	0
4	Berczy Park	8	2	0

Analysis and K-Means Clustering:

After preprocessing the data, we are ready to proceed with clustering the data. In the first step of the process we remove the neighborhoods with one or more 'Sport Goods Shop' from the data frame. The resulting data frame will have 78 neighborhoods. A snapshot of the first few rows of this data frame is shown below.

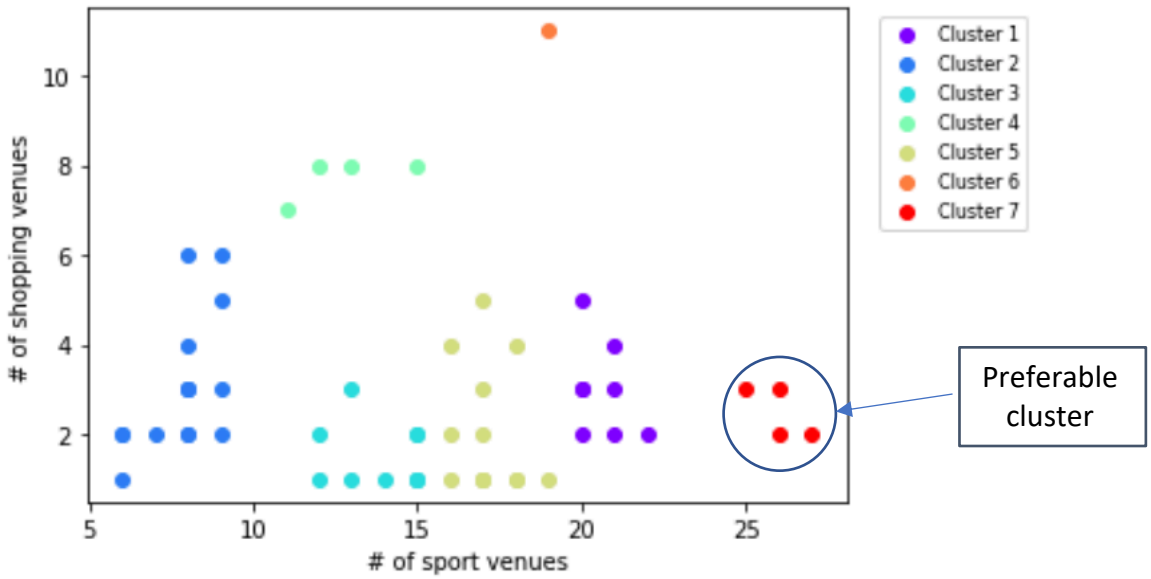
	Neighborhood	# of sport venues	# of shopping venues
1	Alderwood, Long Branch	17	3
2	Bathurst Manor, Wilson Heights, Downsview North	13	1
3	Bayview Village	8	4
4	Berczy Park	8	2
5	Birch Cliff, Cliffside West	17	1

Now we proceed with clustering the data. The following figure visualizes the resulting clusters on the Toronto map.



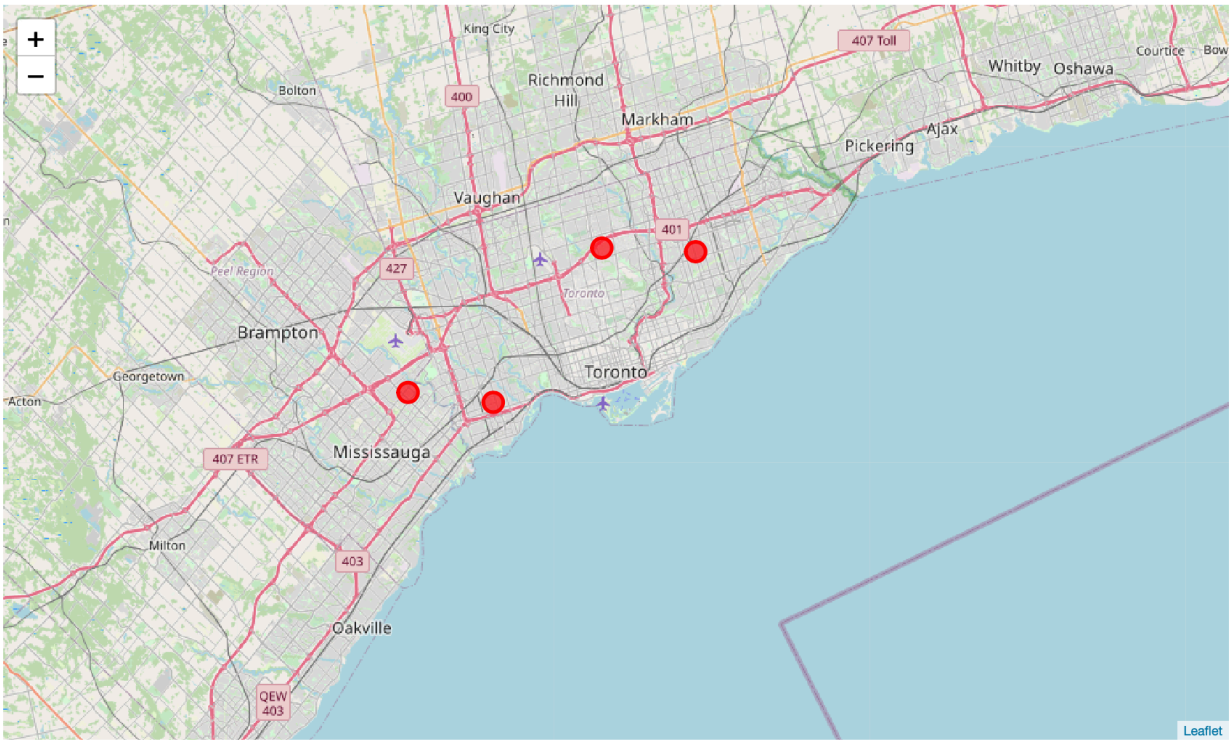
Results and discussion:

To make a better understanding of the clusters and make better judgment of the results, we develop the following scatter plot which shows the number of shopping venues (available shopping plazas or shopping malls to open a sporting goods shop) against the number of sport venues for each neighborhood. Different clusters in this figure are shown with different colors.



From this figure, it seems that neighborhoods in **Cluster 7** are superior to other neighborhoods because compared to other neighborhoods, these neighborhoods have more sport venues while they have at least 2 shopping plaza or shopping mall to open a **Sporting Goods Shop**. The following figures shows the details of this neighborhoods and their locations on the Toronto Map.

	Cluster Labels	Neighborhood	# of sport venues	# of shopping venues	Postalcode	Borough	Latitude	Longitude
7	6	Canada Post Gateway Processing Centre	26	2	M7R	Mississauga	43.636966	-79.615819
42	6	Mimico NW, The Queensway West, South of Bloor,...	26	3	M8Z	Etobicoke	43.628841	-79.520999
74	6	Wexford, Maryvale	25	3	M1R	Scarborough	43.750071	-79.295849
78	6	York Mills West	27	2	M2P	North York	43.752758	-79.400049



In this study we explored different neighborhoods of Toronto and used data to find the potential locations for opening a sporting goods shop. We used data collection tools such as Foursquare API, descriptive data analysis, data visualization, and machine learning techniques such as K-Means clustering to solve our problem. Form the results we could see that we can use data science tools to help solve problems and make more informed decisions.