Coursera Capstone - The Battle of Neighborhoods Report

1. Introduction/Business Problem

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2,731,571 in 2016. Current to 2016, the Toronto census metropolitan area (CMA), of which the majority is within the Greater Toronto Area (GTA), held a population of 5,928,040, making it Canada's most populous CMA. Toronto 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 an international centre for business and finance. Generally considered the financial capital of Canada, Toronto has a high concentration of banks and brokerage firms on Bay Street, in the Financial District. The Toronto Stock Exchange is the world's seventh-largest stock exchange by market capitalization. The five largest financial institutions of Canada, collectively known as the Big Five, have national offices in Toronto.

Toronto's unemployment rate was 6.7% as of July 2016. According to the website Numbeo, Toronto's cost of living plus rent index was second highest in Canada (of 31 cities). The local purchasing power was the sixth lowest in Canada, mid-2017. The average monthly social assistance caseload for January to October 2014 was 92,771. The number of seniors living in poverty increased from 10.5% in 2011 to 12.1% in 2014.

The city's population grew by 4% (96,073 residents) between 1996 and 2001, 1% (21,787 residents) between 2001 and 2006, 4.3% (111,779 residents) between 2006 and 2011, and 4.5% (116,511) between 2011 and 2016. In 2016, persons aged 14 years and under made up 14.5% of the population, and those aged 65 years and over made up 15.6%. The median age was 39.3 years. The city's gender population is 48% male and 52% female. Women outnumber men in all age groups 15 and older.

For those facts economic and demographic I am wondering if open a business in Toronto could be a good idea for investing in the area of fitness. In this case I am looking for Gyms in Toronto Boroughs, for that reason I will do a research using geo-location data and the Foursquare API to find an optimum location in the city of Toronto to open the business and minimize the risk of failure.

2. Data

As our idea of business is open Gym it will be used the Foursquare API to find the gym frequency in a neighborhood or existence of a trending gym are another data that can be useful in the project.

We will then be able to come up with the best location for the gym with all these features, techniques and data. The location is the optimum neighborhood to start offering services. For example, a neighborhood with a lot of available gyms or a trending high-end gym, will be classified as a high risk. The model intended to recommend a neighborhood where will be a higher demand of gym service due to the absence of gyms in that area.

Toronto neighborhood/borough data set

	Postcode	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront
3	M5A	Downtown Toronto	Regent Park
4	M6A	North York	Lawrence Heights

Source: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Demographics of Toronto neighborhoods data set

	Neighbourhood	Population	Land Area	Density	Population Change	Average Income	Transit Commuting	2nd Language	2nd Language %
0	Agincourt	44,577	12.45	3580	4.6	25,750	11.1	Cantonese (19.3%)	19.3% Cantonese
1	Alderwood	11,656	4.94	2360	-4.0	35,239	8.8	Polish (6.2%)	06.2% Polish
2	Alexandra Park	4,355	0.32	13,609	0.0	19,687	13.8	Cantonese (17.9%)	17.9% Cantonese
3	Allenby	2,513	0.58	4333	-1.0	245,592	5.2	Russian (1.4%)	01.4% Russian
4	Amesbury	17,318	3.51	4,934	1.1	27,546	16.4	Spanish (6.1%)	06.1% Spanish

Source: https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

Toronto gym data

Foursquare is a local search-and-discovery service mobile app which provides search results for its users. The app provides personalized recommendations of places to go to near a user's current location based on users' "previous browsing history, purchases, or check-in history".

Foursquare API will be used to explore the various types of venues and their categories available in each neighborhood.

Source: https://developer.foursquare.com/

Geospatial Coordinates

Geospatial coordinates are used to complete the neighborhood data with missing latitude and longitude. Those latitude and longitude data are used for k-means clustering and visualizing neighborhoods on Toronto Map.

	Postcode	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Source: Geospatial_Coordinates.csv (Used in the previous courses before)

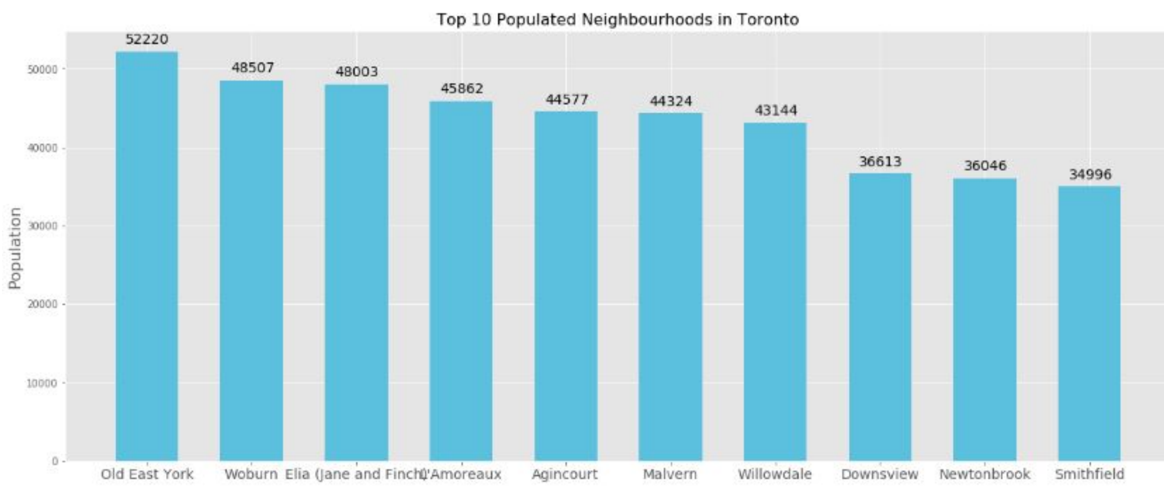
3. Methodology

Preparing the data Toronto neighborhood, demographics and geospatial data merged in order to be handled easily. Population score added to that dataframe which is the percentage of the population among the Toronto population. After that, the missing latitude and longitude data are found with the geopy.geocoders and inserted to table.

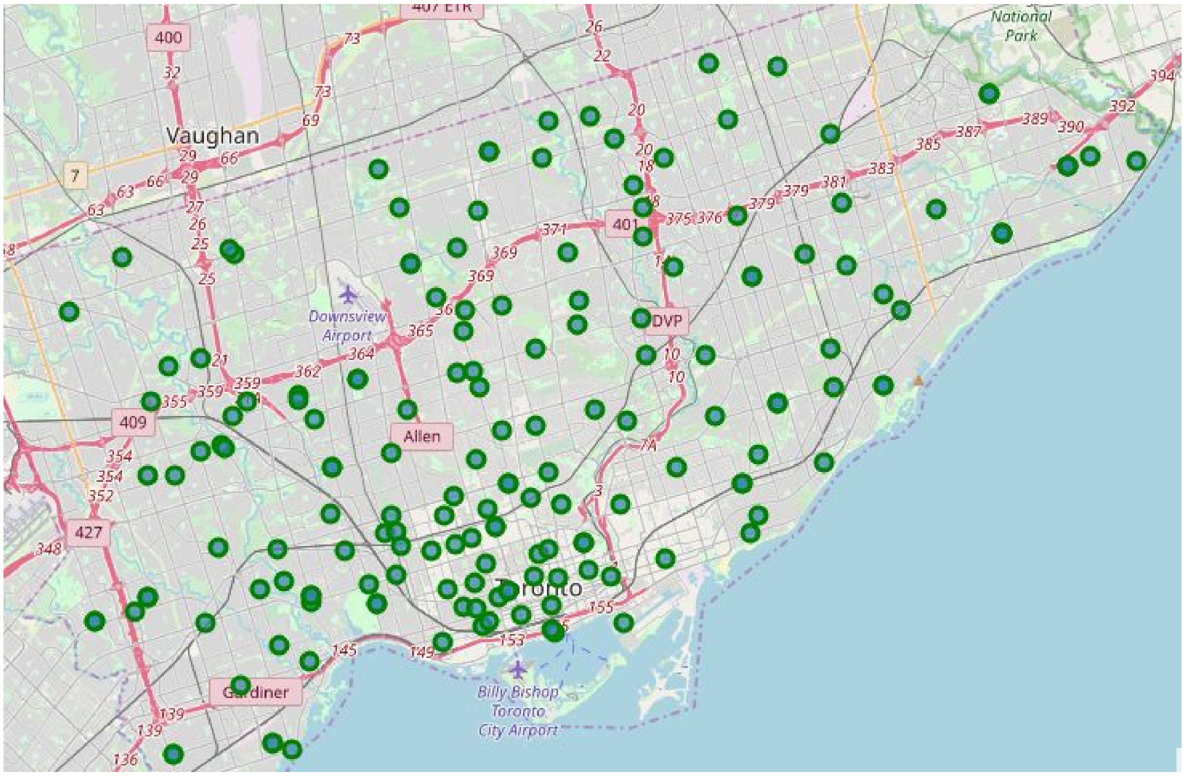
Toronto neighborhoods data after cleaning and processing

	Neighbourhood	Population	Land Area	Density	Population Change	Average Income	Transit Commuting	2nd Language	2nd Language %	Borough	Postcode	Latitude
0	Agincourt	44577	12.45	3580	4.6	25,750	11.1	Cantonese (19.3%)	19.3% Cantonese	Scarborough	M1S	43.7942
1	Alderwood	11656	4.94	2360	-4.0	35,239	8.8	Polish (6.2%)	06.2% Polish	Etobicoke	M8W	43.6024
2	Alexandra Park	4355	0.32	13,609	0.0	19,687	13.8	Cantonese (17.9%)	17.9% Cantonese			43.6508
3	Allenby	2513	0.58	4333	-1.0	245,592	5.2	Russian (1.4%)	01.4% Russian			43.7114
4	Amesbury	17318	3.51	4,934	1.1	27,546	16.4	Spanish (6.1%)	06.1% Spanish			43.7062

Top 10 Neighborhoods in Toronto by Population



Neighborhoods on the data visualized on Toronto Map



Finding the gyms in every neighborhood with foursquare API

Search queries formed for every neighborhood in the data set in order to retrieve gyms in them. 164 API requests are sent and 334 venues found. After dropping non-gym venues and duplicates, there are 124 gyms left in Toronto.

Gym data after cleaning and processing

	Name	Neighbourhood	Category	Distance	Latitude	Longitude	VenueID
0	Sheraton Gateway Gym	Allenby	Gym / Fitness Center	6040	43.686321	-79.620017	4f8cecf4e4b04bd7c548047f
1	Gym	Bayview Woods – Steeles	Gym	5975	43.842250	-79.425346	51caef5a498ed37e7db65cf1
2	Bayview Place Gym	Bayview Woods – Steeles	Gym	5087	43.841229	-79.404016	51f53565498e8378e402aaea
3	Gym @ Vista/Beverly Condo	Branson	Gym	4627	43.812821	-79.452578	4d5a90ab35966dcbaa786228
4	West Harbour City Gym	Brockton	Gym	2034	43.636440	-79.402944	4d690bf8b6f46dcb357d1cb2

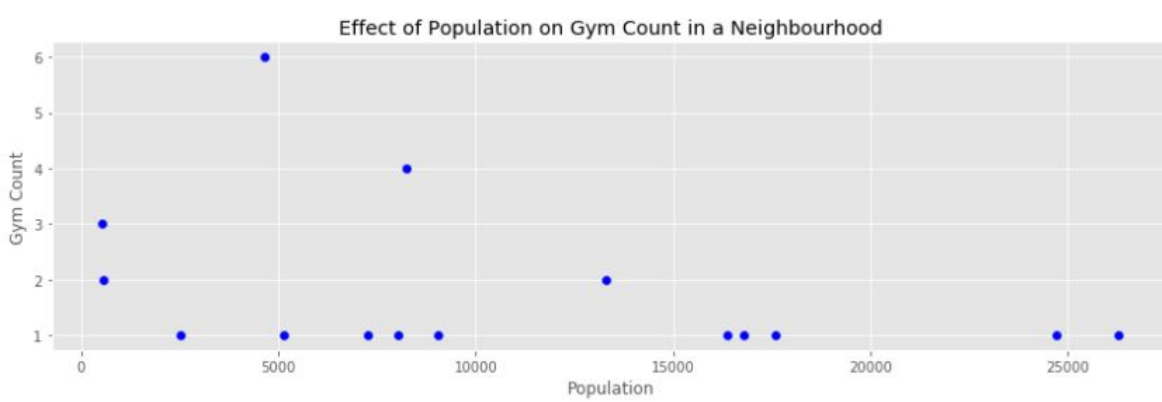
Gym data grouped by neighborhoods and Fashion District has the most gyms in Toronto (Weight is the percentage of gyms within the total, e.g. Fashion District has the %19.35 of the gyms in Toronto.)

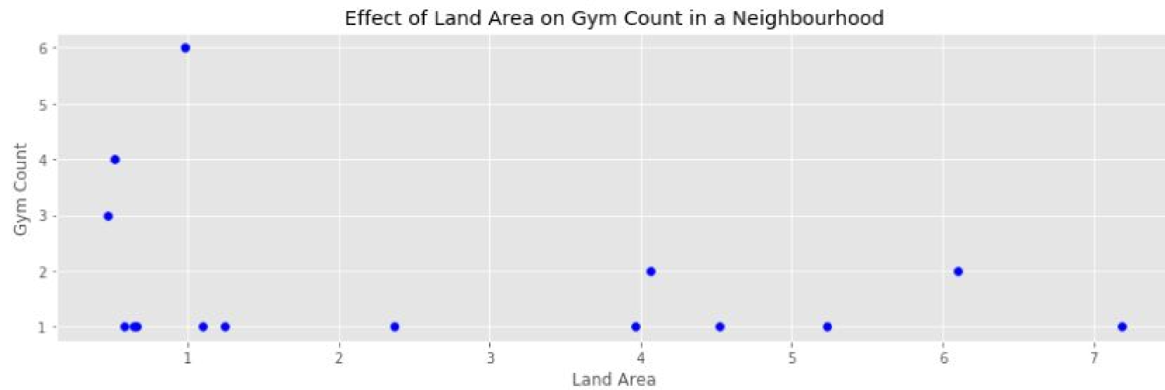
Neighbourhood	GymCount
Fashion District	6
Garden District	4
Financial District	3
Bayview Woods – Steeles	2
Port Lands	2
Allenby	1
Branson	1
Brockton	1
Christie Pits	1
Discovery District	1

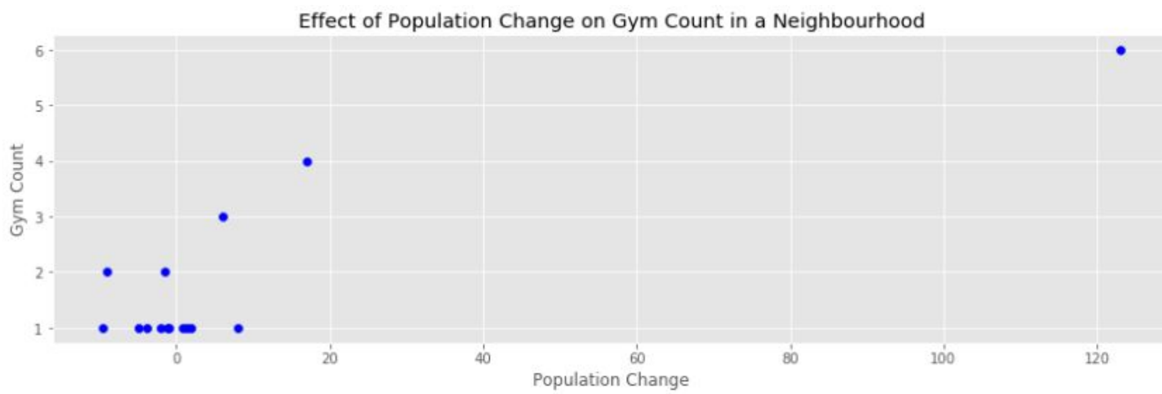
Visualizing the data

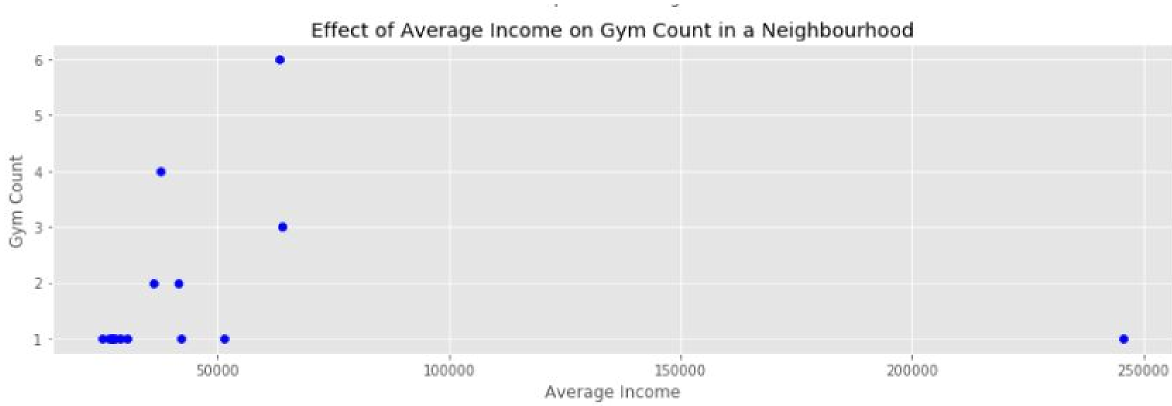
Before jumping into machine learning, the data is visualized in order to find the most suitable machine learning technique. The gym count in a neighborhood is taken as the dependent variable and other variables are taken as the independent variables in those scatter plots. Those plots clearly show that there is no significant linear relationship between gym count and those variables.

Population, Land Area, Population Change, Average Income / Gym Count in a Neighborhood scatter plots









In those plots, there were multiple extreme outliers and not a significant linear relationship was encountered. Correlation of those variables confirmed this belief.

Correlation table of population, land area, population change, average income and gym count

	population	land_area	population_change	average_income	gymcount
population	1.000000	0.637577	-0.197208	-0.381050	-0.391859
land_area	0.637577	1.000000	-0.263936	-0.302986	-0.295886
population_change	-0.197208	-0.263936	1.000000	0.061153	0.848735
average_income	-0.381050	-0.302986	0.061153	1.000000	0.012928
gymcount	-0.391859	-0.295886	0.848735	0.012928	1.000000

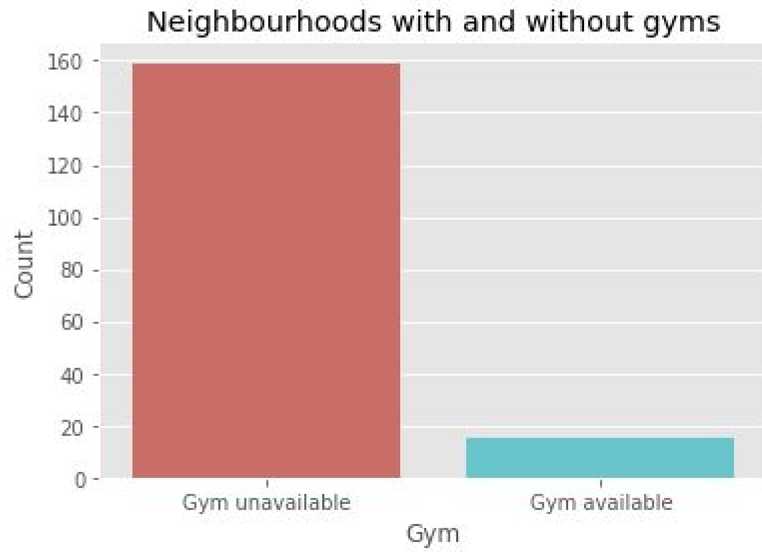
Logistic Regression

Since the variables doesn't have a linear relationship between them, the gym data is converted to one hot encoding in order to apply logistic regression.

Gym one hot encoding(Neighborhoods that have at least 1 gym have 1 at gym column, otherwise 0 at gym column)

	neighbourhood	population	land_area	population_change	average_income	borough	postcode	latitude	longitude	population_sc
0	Agincourt	44577	12.45	4.6	25750	Scarborough	M1S	43.7942	-79.262	1.845
1	Alderwood	11656	4.94	-4.0	35239	Etobicoke	M8W	43.6024	-79.5435	0.482
2	Alexandra Park	4355	0.32	0.0	19687			43.6508	-79.4043	0.180
3	Allenby	2513	0.58	-1.0	245592			43.7114	-79.5534	0.104
4	Amesbury	17318	3.51	1.1	27546			43.7062	-79.4835	0.716

Distribution of neighborhoods with and without a gym



From this data, we expect to find neighborhoods that share same characteristics and features. The accuracy of the model was high and it was predicting the class 0 (No gym) with a high probability. However, the model wasn't able to predict class 1 (Gym) as good as class 0.

Score of the logistic regression model.

```
Accuracy of logistic regression classifier on test set: 0.92
```

Probability of class 0 and class 1 in the first five neighborhoods. (e.g. The first neighborhood has no gym by 72% chance and it has a gym by 28% chance.)

```
array([[0.8968445 , 0.1031555 ],
       [0.85877909, 0.14122091],
       [0.89523236, 0.10476764],
       [0.89541706, 0.10458294],
       [0.92047293, 0.07952707]])
```

The confusion matrix of the logistic regression model on a training set looked like this. It predicted 2 out of 9 neighborhoods with a gym correctly and 44 out of 44 neighborhoods without a gym correctly

```
array([[ 0,  3],
       [ 1, 49]])
```

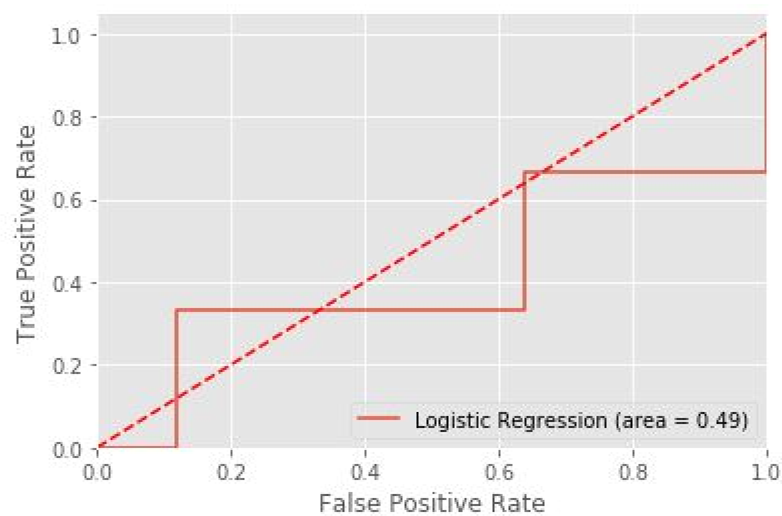
The classification report of the model. Even though the class 1 has bad results, the class 0 saves the model.

```
              precision    recall  f1-score   support

           0       0.94      0.98      0.96        50
           1       0.00      0.00      0.00         3

   micro avg       0.92      0.92      0.92        53
   macro avg       0.47      0.49      0.48        53
weighted avg       0.89      0.92      0.91        53
```

The roc curve of the model.



The model was good enough to apply to the whole set in order find neighborhoods with same features. Model predicted those neighborhoods with gyms and it was correct most of the time and they have at least one gym. However, only the Rouge neighborhood predicted as having a gym, even though it doesn't have one.

Neighborhoods with a gym according to the logistic regression model.

	neighbourhood	latitude	longitude	gymcount	gymscore	gym	Predicted Values	Prediction Probability
130	Rouge	43.8067	-79.1944	0	0.0	0	1	0.524804

K-Means Clustering

K-Means clustering will be another technique to cluster neighborhoods with shared characteristics and features. The previous neighborhood data is clustered with k=15 and fixed random_state=150 for the best results. Cluster distribution is moderately balanced and there is no bias in terms of gym count.

Cluster labels of every single neighborhood.

```
array([ 5,  0, 12,  9, 10,  1,  3,  0,  8,  7,  8,  3,  6,  7, 13,  8, 12,
        2, 12,  7, 11, 12,  3,  0,  3,  0,  0,  0,  0,  0, 13, 12, 11, 12,
        7,  3,  8,  7,  6, 10, 12,  5, 12, 10, 10, 10, 10,  5,  0,  6, 13,
       13, 10, 11,  8,  0, 10,  0, 12,  8,  8, 13, 12, 13,  7,  0, 10,  4,
        0,  8, 12,  3,  0, 12, 12,  7, 12, 12, 10,  0,  5,  8, 12,  0,  4,
        3,  3,  6,  0, 12,  0,  1,  5, 10,  7,  0,  6,  7, 14, 12, 10,  0,
        5,  8,  0, 12, 10,  5, 11,  6, 12,  6,  0, 10,  8,  7,  3, 12,  6,
        6,  7,  4,  6,  0,  8,  6, 10, 12,  8, 10,  5,  1, 12,  6,  3,  7,
       13,  6, 13, 13,  8, 12,  0, 11,  8, 10, 10,  8,  0,  7, 10, 12,  8,
        6,  8, 10,  0, 10, 10,  5,  0,  5, 13,  3,  6, 11], dtype=int32)
```

Neighborhood counts in each cluster.

	neighbourhood
cluster	
0	27
12	25
10	22
8	18
6	15
7	13
3	11
5	10
13	10
11	6
1	3
4	3
2	1
9	1
14	1

After the clustering, the labels and cluster score are added to the data set. The cluster score is basically the gym count of the cluster divided by the neighborhood count of the cluster. This is used to represent the best cluster in terms of likeliness of gym count. Cluster 13 has the best score out of 15 clusters. This means cluster 13 is the best cluster in terms of gym count.

Cluster ranking by gym counts / neighborhood counts

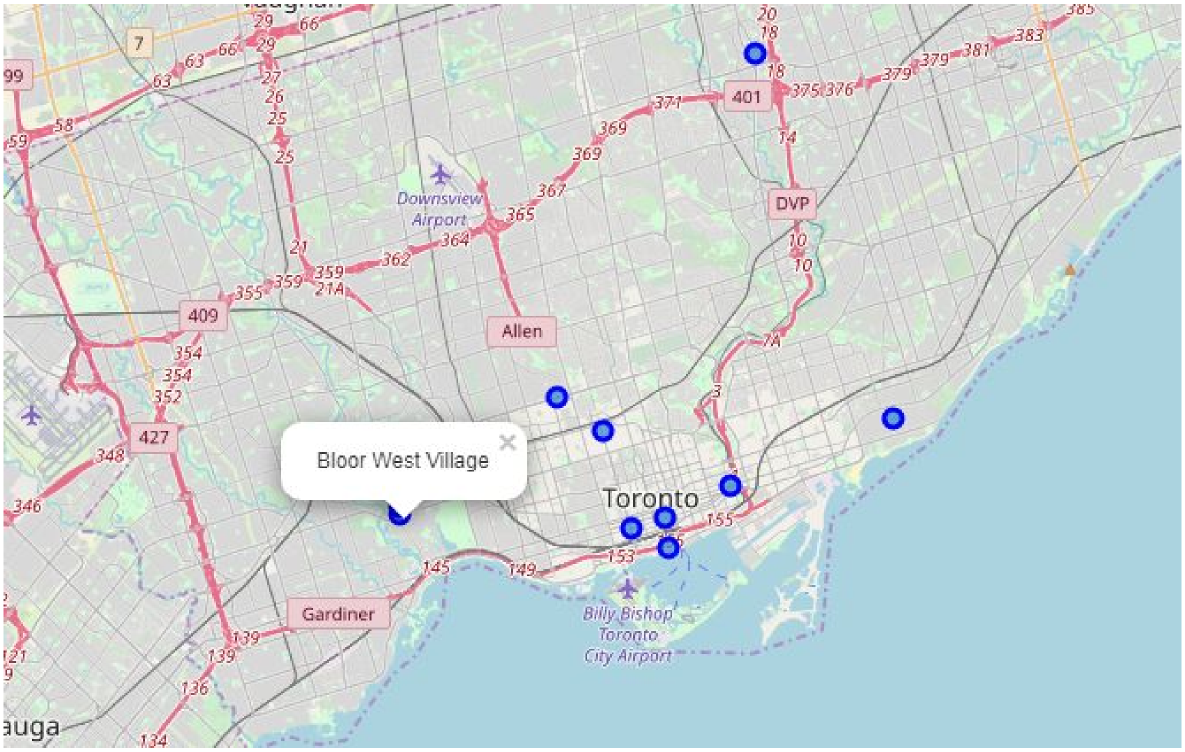
cluster	population	land_area	population_change	average_income	population_score	gymcount	gymscore	gym	score
9	2513	0.58	-1.0	245592	0.104025	1	3.703704	1	3.703704
13	5190	1.45	129.0	127234	0.214838	9	33.333333	2	3.333333
8	21131	10.83	-2.5	119726	0.874709	5	18.518519	3	1.028807
0	13364	1.16	12.0	68170	0.553197	5	18.518519	2	0.685871
6	50968	11.71	-0.1	51903	2.109800	2	7.407407	2	0.493827
10	33176	6.34	2.2	56781	1.373307	2	7.407407	2	0.336700
12	17056	2.35	-13.3	54416	0.706026	2	7.407407	2	0.296296
7	17602	5.23	-1.1	51398	0.728628	1	3.703704	1	0.284900

The best cluster has 6 neighborhoods. 5 of them have multiple gyms except Bloor West Village. Only Bloor West Village doesn't have a gym, but still it is clustered with those neighborhoods, so they share other features.

The best cluster in terms of gym count / neighborhood count

	neighbourhood	population	land_area	population_change	average_income	latitude	longitude	population_score	gymcount
14	Bloor West Village	5175	0.74	-2.0	55578	43.6493	-79.4844	0.214217	0
31	Corktown	4484	0.67	77.0	54681	43.6574	-79.3565	0.185613	0
52	Fashion District	4642	0.98	123.0	63282	43.6455	-79.395	0.192154	6
53	Financial District	548	0.47	6.0	63952	43.6487	-79.3815	0.022684	3
64	Harbourfront / CityPlace	14368	1.87	94.3	69232	43.6401	-79.3801	0.594758	0
66	Henry Farm	2790	0.91	-6.0	56395	43.7785	-79.3466	0.115491	0
144	Swansea	11133	3.76	0.5	58681	43.6516	-79.4844	0.460846	0
146	The Annex	15602	1.47	-2.3	63636	43.6727	-79.4057	0.645839	0
147	The Beaches	20416	3.57	7.8	67536	43.6764	-79.293	0.845112	0

The best cluster visualized on Toronto map



4. Result

Logistic Regression

The logistic regression model was pretty good at predicting neighborhoods without a gym, but it was struggling at predicting neighborhoods with gym. It classified those neighborhoods with a gym and it was right.

However, Rouge was predicted as a neighborhood with a gym, even though it doesn't have one. This means it has the same features with other neighborhoods that has multiple gyms and we can assume that a gym can be successful in that neighborhood. Class 1 predictions by the logistic regression model

K-Means Clustering

The k-means clustering model was good at clustering neighbourhoods with high number of gyms. The best cluster was selected among the clusters in terms of gym count divided by neighborhood count. The best cluster has neighborhoods with multiple gyms, but Bloor West Village doesn't have any gym. It shares similarities with other neighborhood, so a gym in that neighborhood can be successful.

According to two different models, Bloor West Village and Willowdale are the most similar with other neighborhoods that has at least one gym. Selecting either one of those will lower the risk to minimum because the similar neighborhoods have a higher demand of gyms and those two doesn't have a gym.

	neighbourhood	population	land_area	population_change	average_income	borough	postcode	latitude	longitude	population_scoı
14	Bloor West Village	5175	0.74	-2.0	55578			43.6493	-79.4844	0.21421
167	Willowdale	43144	7.68	62.3	39895	North York	M2M	43.7891	-79.4085	1.78592

5. Discussion

The models shown above was good at classifying the neighborhoods with similar features, but they were not perfect. That's why they can't predict everything correctly and they shouldn't. For instance the logistic regression model had only one neighborhood which is classified as having a gym even though it doesn't have one. It was a mistake but it made me think that neighborhood should have a gym because other neighborhoods are very similar to that neighborhood in terms average income, population, land area etc. and they have a gym.

The k-means model made a cluster with neighborhoods which has multiple gyms after many attempts of different k's and random states. That cluster had neighborhoods with highest number of gyms and a single neighborhood without a gym (Bloor West Village). It also looks like a mistake but the points stated above is valid for this model as well. I selected those two neighborhoods because they don't have any gyms at all. There could be any neighborhood which is more similar to other neighborhoods with high number of gyms than those two selected neighborhoods. However, these two were the only ones without a gym and starting the business in those neighborhoods would give competitive advantage unlike other neighborhoods.

6. Conclusion

To conclude the best neighborhood recommendations for starting a gym are Willowdale and Bloor West Village. The key factors for selecting those neighborhoods are likeliness with other neighborhoods which has higher demand for gyms. Their likeliness comes from factors such as population, land area, population change, average income, coordinates, etc.

This project can be replicated for any type of business in any location. The project doesn't imply that starting a gym in those neighborhood will be successful no matter what. The project shows that those two neighborhoods are very similar to other neighborhoods with multiple gyms, so the demand will be similar as well.