Coursera Capstone - The Battle of Neighborhoods Report

1) Introduction

Toronto is the largest city in Canada by population, with 2,731,571 residents in 2016. A global city, Toronto is a center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

Toronto has teams in nearly every major professional sport, including the Toronto Blue Jays (MLB), Toronto Argonauts (CFL), Toronto Raptors (NBA), Toronto Rock (NLL), Toronto FC (MLS), Ontario Arrows (MLR) and the Toronto Maple Leafs (NHL). This shows there is a strong sports culture in Toronto.

Its economy is highly diversified with strengths in technology, design, financial services, life sciences, education, arts, fashion, business services, environmental innovation, food services, and tourism. That means the market is highly competitive for people who want to start business there. A new business plan in Toronto needs to be designed carefully, so that the business will be successful and sustainable. The business plan's aim should be lowering the risk to minimum.

- **The business problem**

I want to open a gym in Toronto and I have to resolve the problem stated above in order to be successful. Otherwise, the business will be a failure. The first location of the gym needs to be chosen carefully because of the points highlighted above. The first gym's success is more important than the later ones because it determines the fate of the business mostly. If it is successful, the same success can be achieved incrementally with second and third gyms. That's why the first gym and its location (i.e. neighborhood) is crucial for the business. In order to solve this problem, I am going to find the optimum location for a gym in the city of Toronto and it will minimize the failure risk related to the location

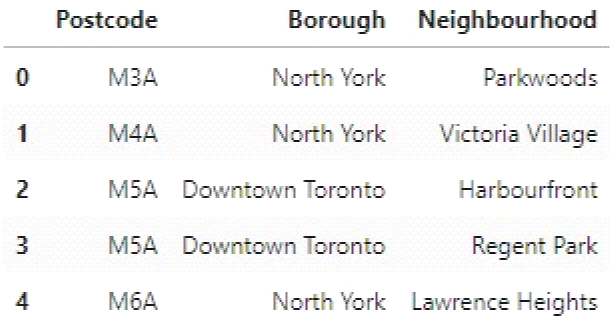
- **Who would be interested in this project?**

Anyone who wants to start a gym in Toronto can benefit from this project. In addition to that, anyone who has a similar problem can replicate the data analysis and machine learning techniques used in this project to solve their problem.

2) Data

- **Toronto neighborhood/borough data set**

The Toronto neighborhood data set is going to be used for segmenting and clustering the neighborhoods in the city of Toronto. It will help to group neighborhoods and boroughs of Toronto. The data set contains borough and postcode of each neighborhood in Toronto.

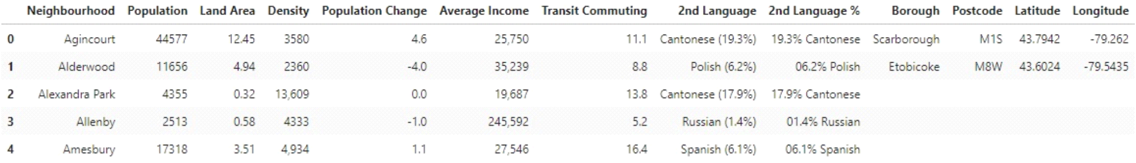


	Postcode	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront
3	M5A	Downtown Toronto	Regent Park
4	M6A	North York	Lawrence Heights

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

- **Demographics of Toronto neighborhoods data set**

The demographics of Toronto data set will help to detect the similarities and dissimilarities of neighborhoods. This data set has sufficient data for applying techniques such as logistic regression and k-means clustering. The data set contains the coordinates for each of the neighborhood in Toronto. The data set contains useful information such as census tracts, population, land area, density, population change, average income, transit commuting percentage, renters percentage, second most common language by name, second most common language by percentage, etc.

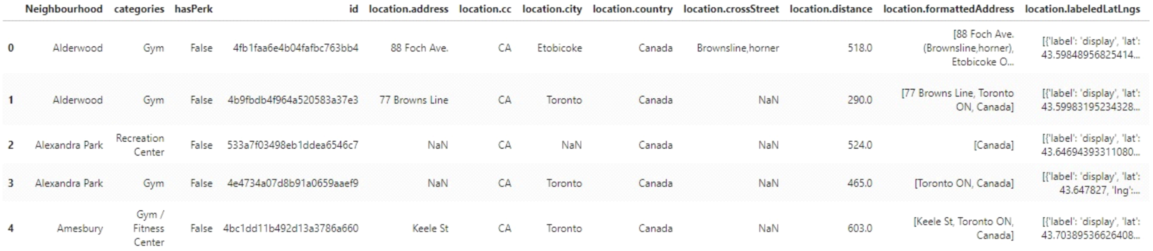


	Neighbourhood	Population	Land Area	Density	Population Change	Average Income	Transit Commuting	2nd Language	2nd Language %	Borough	Postcode	Latitude	Longitude
0	Agincourt	44577	12.45	3580	4.6	25,750	11.1	Cantonese (19.3%)	19.3% Cantonese	Scarborough	M1S	43.7942	-79.262
1	Alderwood	11656	4.94	2360	-4.0	35,239	8.8	Polish (6.2%)	06.2% Polish	Etobicoke	M8W	43.6024	-79.5435
2	Alexandra Park	4355	0.32	13,609	0.0	19,687	13.8	Cantonese (17.9%)	17.9% Cantonese				
3	Allenby	2513	0.58	4333	-1.0	245,592	5.2	Russian (1.4%)	01.4% Russian				
4	Amesbury	17318	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 gyms available in each neighborhood. The trending venues in a neighborhood can be displayed with the API. The gym frequency in a neighborhood or existence of a trending gym can be used in the project.

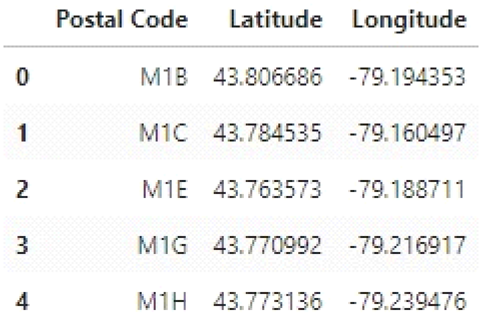


	Neighbourhood	categories	hasPerk	id	location.address	location.cc	location.city	location.country	location.crossStreet	location.distance	location.formattedAddress	location.labeledLatLngs
0	Alderwood	Gym	False	4fb1faa6e4b04fafbc763bb4	88 Foch Ave.	CA	Etobicoke	Canada	Brownsline,horner	518.0	[88 Foch Ave. (Brownsline,horner), Etobicoke O...	[{'label': 'display', 'lat': 43.59848956825414...
1	Alderwood	Gym	False	4b9fbdb4f964a520583a37e3	77 Browns Line	CA	Toronto	Canada	NaN	290.0	[77 Browns Line, Toronto ON, Canada]	[{'label': 'display', 'lat': 43.59983195234328...
2	Alexandra Park	Recreation Center	False	533a7f03498eb1ddea6546c7	NaN	CA	NaN	Canada	NaN	524.0	[Canada]	[{'label': 'display', 'lat': 43.64694393311080...
3	Alexandra Park	Gym	False	4e4734a07d8b91a0659aaef9	NaN	CA	Toronto	Canada	NaN	465.0	[Toronto ON, Canada]	[{'label': 'display', 'lat': 43.647827, 'lng':...
4	Amesbury	Gym / Fitness Center	False	4bc1dd11b492d13a3786a660	Keele St	CA	Toronto	Canada	NaN	603.0	[Keele St, Toronto ON, Canada]	[{'label': 'display', 'lat': 43.70389536626408...

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.



	Postal Code	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)

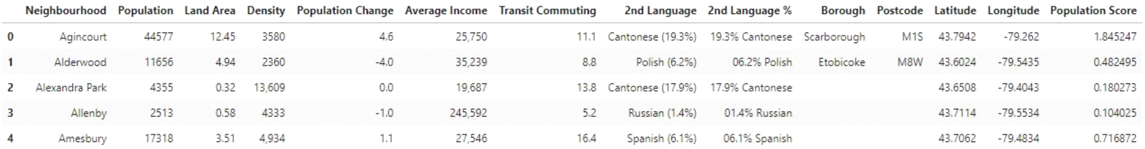
3) Methodology

- **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.

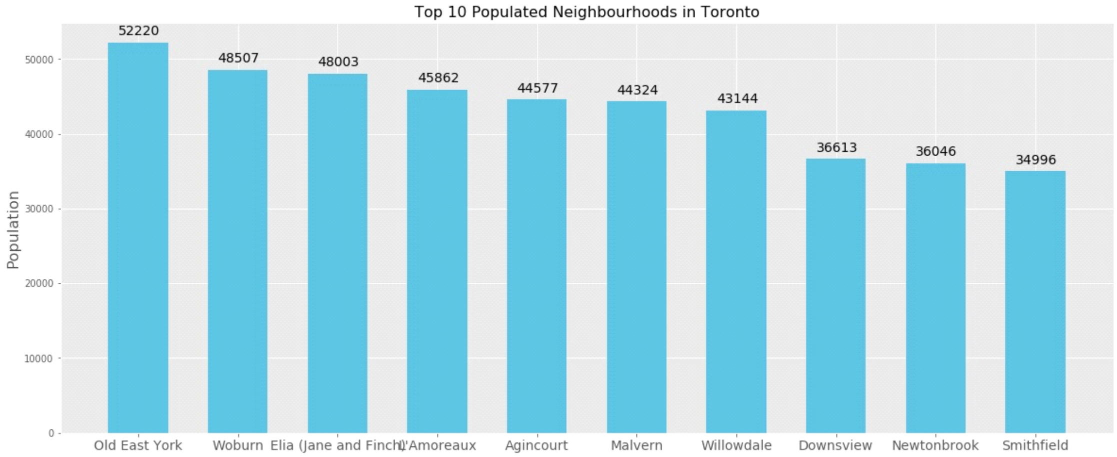
After that, the missing latitude and longitude data are found with the geopy.geocoders and inserted to table.

Toronto neighborhoods data after cleaning and processing

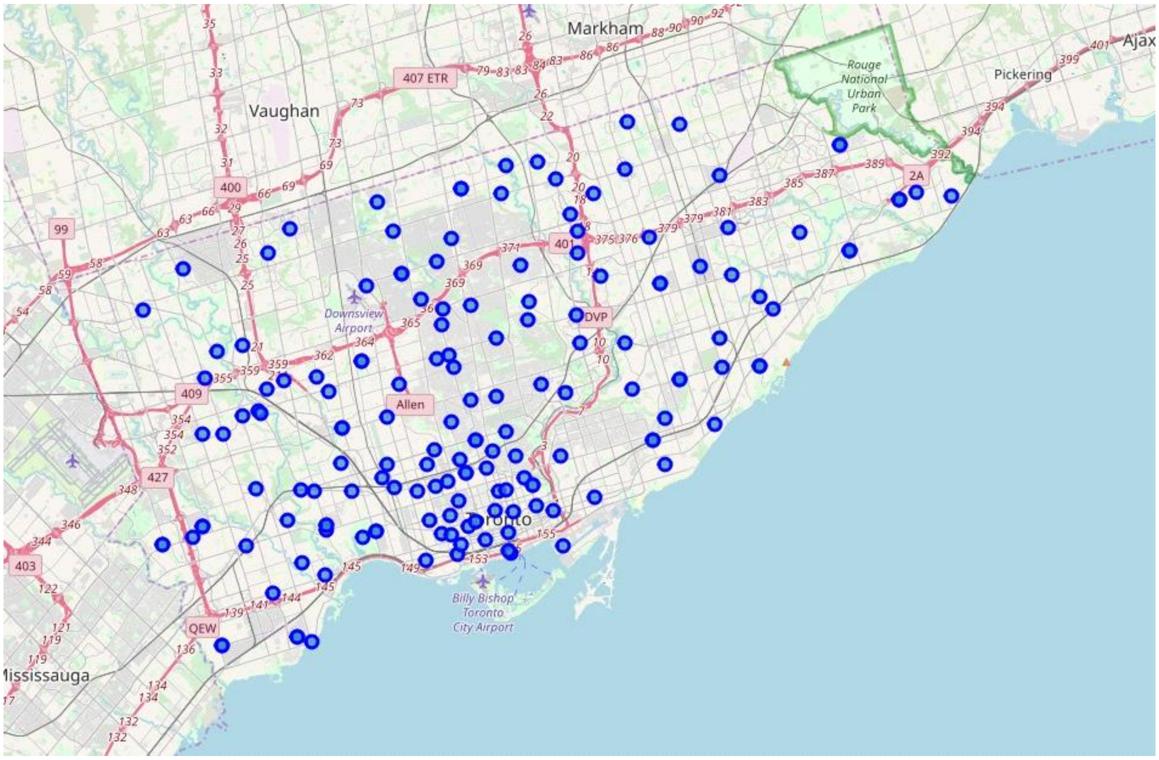


	Neighbourhood	Population	Land Area	Density	Population Change	Average Income	Transit Commuting	2nd Language	2nd Language %	Borough	Postcode	Latitude	Longitude	Population Score
0	Agincourt	44577	12.45	3580	4.6	25,750	11.1	Cantonese (19.3%)	19.3% Cantonese	Scarborough	M1S	43.7942	-79.262	1.845247
1	Alderwood	11656	4.94	2360	-4.0	35,239	8.8	Polish (6.2%)	06.2% Polish	Etobicoke	M8W	43.6024	-79.5435	0.482495
2	Alexandra Park	4355	0.32	13,609	0.0	19,687	13.8	Cantonese (17.9%)	17.9% Cantonese			43.6508	-79.4043	0.180273
3	Allenby	2513	0.58	4333	-1.0	245,592	5.2	Russian (1.4%)	01.4% Russian			43.7114	-79.5534	0.104025
4	Amesbury	17318	3.51	4,934	1.1	27,546	16.4	Spanish (6.1%)	06.1% Spanish			43.7062	-79.4834	0.716872

Top 10 Neighborhoods in Toronto by Population



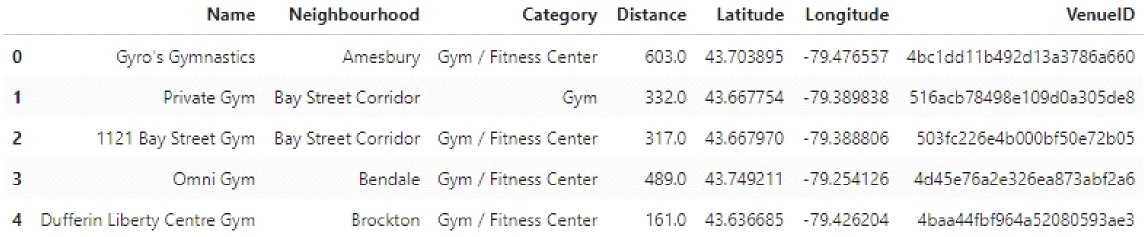
Neighborhoods on the data visualized on Toronto Map



- **Finding the gyms in every neighborhood with foursquare API**

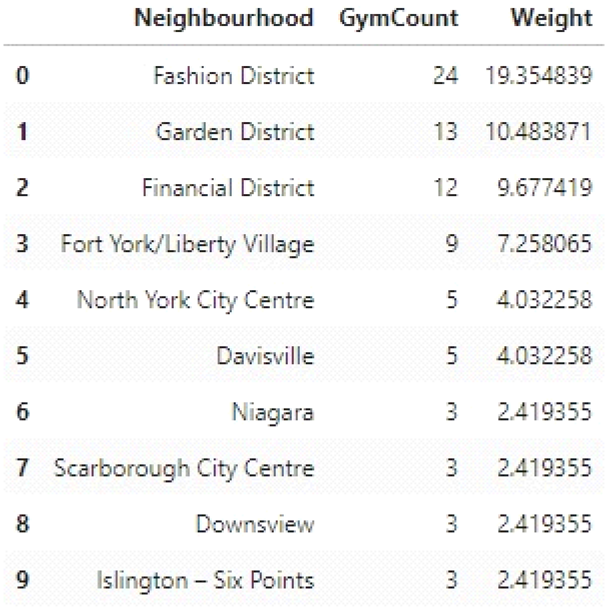
Search queries formed for every neighborhood in the data set in order to retrieve gyms in them. 164 API requests are sent and 334 venues found. After dropping non-gym venues and duplicates, there are 124 gyms left in Toronto.

Gym data after cleaning and processing



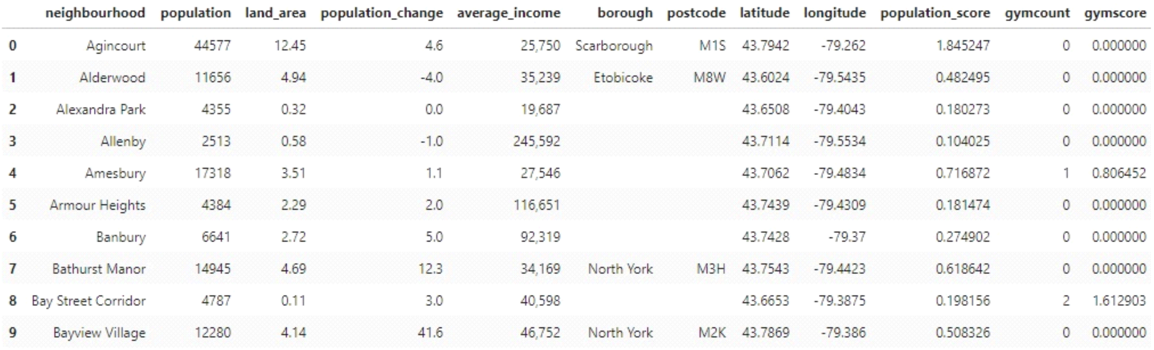
	Name	Neighbourhood	Category	Distance	Latitude	Longitude	VenueID
0	Gyro's Gymnastics	Amesbury	Gym / Fitness Center	603.0	43.703895	-79.476557	4bc1dd11b492d13a3786a660
1	Private Gym	Bay Street Corridor	Gym	332.0	43.667754	-79.389838	516acb78498e109d0a305de8
2	1121 Bay Street Gym	Bay Street Corridor	Gym / Fitness Center	317.0	43.667970	-79.388806	503fc226e4b000bf50e72b05
3	Omni Gym	Bendale	Gym / Fitness Center	489.0	43.749211	-79.254126	4d45e76a2e326ea873abf2a6
4	Dufferin Liberty Centre Gym	Brockton	Gym / Fitness Center	161.0	43.636685	-79.426204	4baa44fbf964a52080593ae3

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	Weight
0	Fashion District	24	19.354839
1	Garden District	13	10.483871
2	Financial District	12	9.677419
3	Fort York/Liberty Village	9	7.258065
4	North York City Centre	5	4.032258
5	Davisville	5	4.032258
6	Niagara	3	2.419355
7	Scarborough City Centre	3	2.419355
8	Downsview	3	2.419355
9	Islington – Six Points	3	2.419355

Gym data merged with neighborhood data. (Weight changed to gymscore)

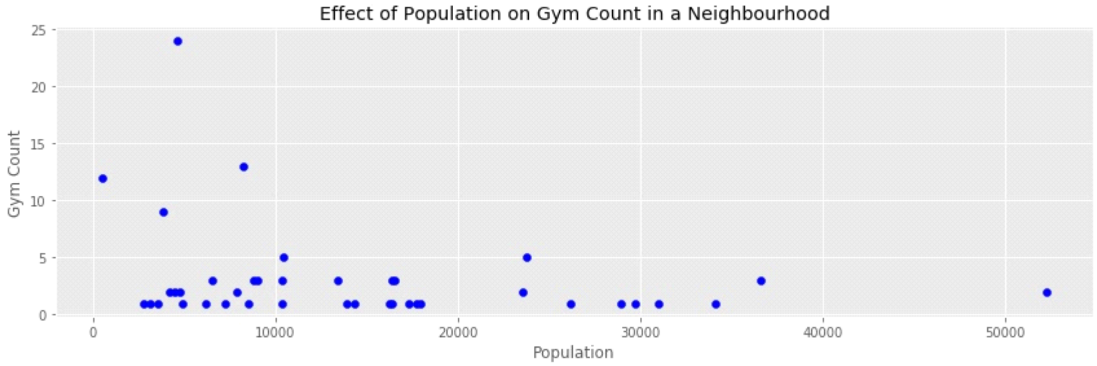


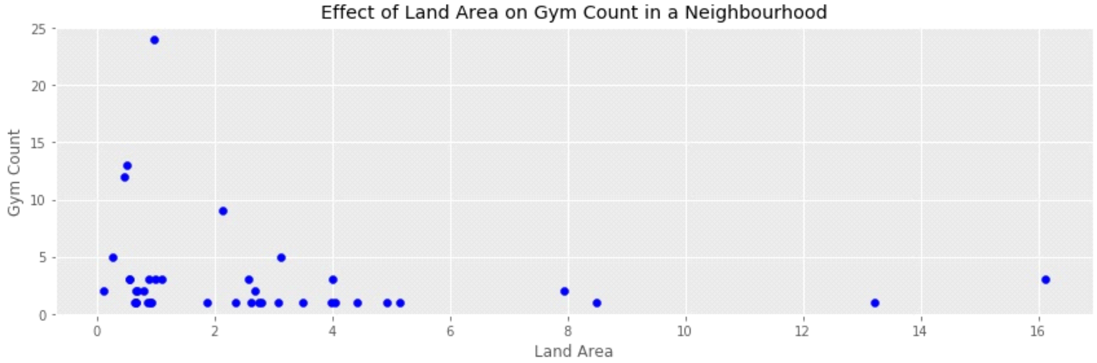
	neighbourhood	population	land_area	population_change	average_income	borough	postcode	latitude	longitude	population_score	gymcount	gymscore
0	Agincourt	44577	12.45	4.6	25,750	Scarborough	M1S	43.7942	-79.262	1.845247	0	0.000000
1	Alderwood	11656	4.94	-4.0	35,239	Etobicoke	M8W	43.6024	-79.5435	0.482495	0	0.000000
2	Alexandra Park	4355	0.32	0.0	19,687			43.6508	-79.4043	0.180273	0	0.000000
3	Allenby	2513	0.58	-1.0	245,592			43.7114	-79.5534	0.104025	0	0.000000
4	Amesbury	17318	3.51	1.1	27,546			43.7062	-79.4834	0.716872	1	0.806452
5	Armour Heights	4384	2.29	2.0	116,651			43.7439	-79.4309	0.181474	0	0.000000
6	Banbury	6641	2.72	5.0	92,319			43.7428	-79.37	0.274902	0	0.000000
7	Bathurst Manor	14945	4.69	12.3	34,169	North York	M3H	43.7543	-79.4423	0.618642	0	0.000000
8	Bay Street Corridor	4787	0.11	3.0	40,598			43.6653	-79.3875	0.198156	2	1.612903
9	Bayview Village	12280	4.14	41.6	46,752	North York	M2K	43.7869	-79.386	0.508326	0	0.000000

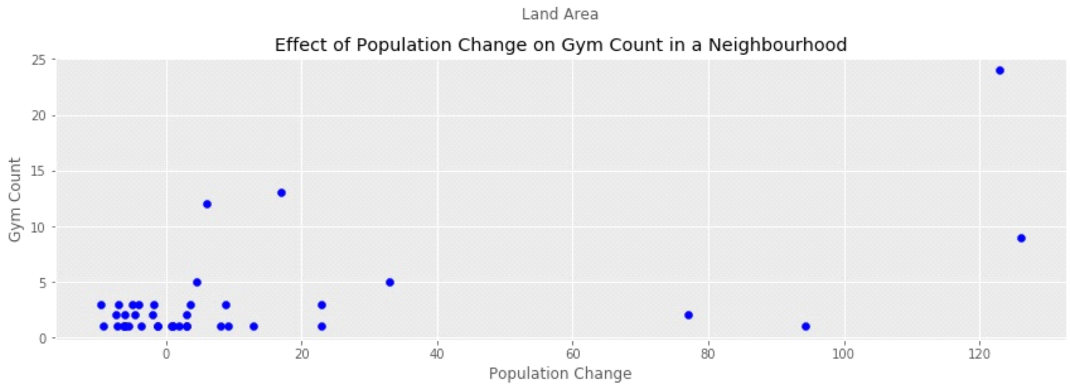
- **Visualizing the data**

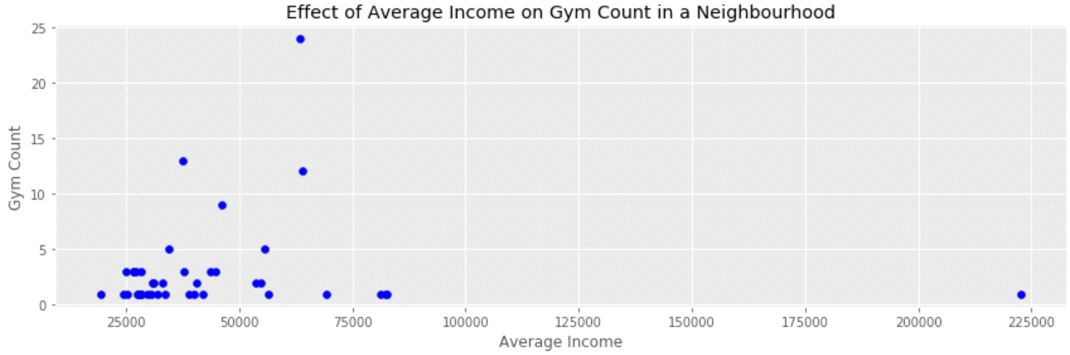
Before jumping into machine learning, the data is visualized in order to find the most suitable machine learning technique. The gym count in a neighborhood is taken as the dependent variable and other variables are taken as the independent variables in those scatter plots. Those plots clearly show that 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


Effect of Population on Gym Count in a Neighbourhood

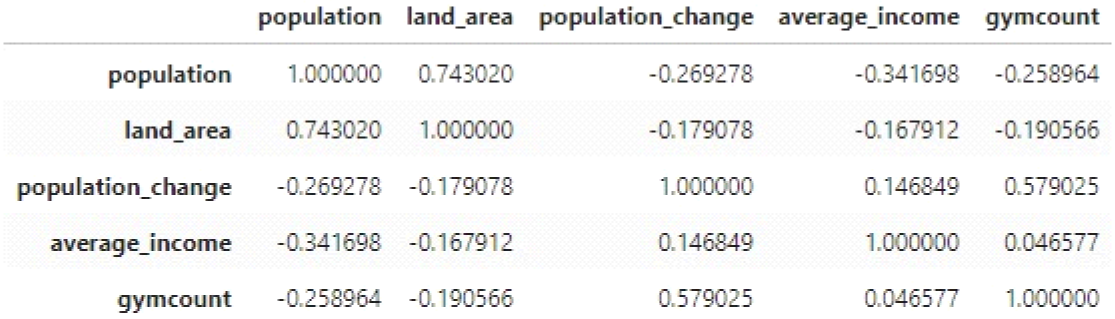

Effect of Land Area on Gym Count in a Neighbourhood


Effect of Population Change on Gym Count in a Neighbourhood


Effect of Average Income on Gym Count in a Neighbourhood

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

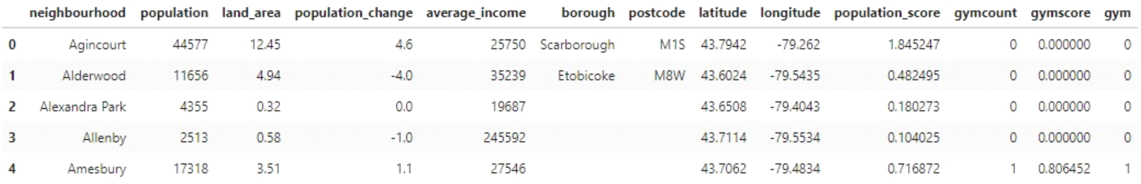


	population	land_area	population_change	average_income	gymcount
population	1.000000	0.743020	-0.269278	-0.341698	-0.258964
land_area	0.743020	1.000000	-0.179078	-0.167912	-0.190566
population_change	-0.269278	-0.179078	1.000000	0.146849	0.579025
average_income	-0.341698	-0.167912	0.146849	1.000000	0.046577
gymcount	-0.258964	-0.190566	0.579025	0.046577	1.000000

- **Logistic Regression**

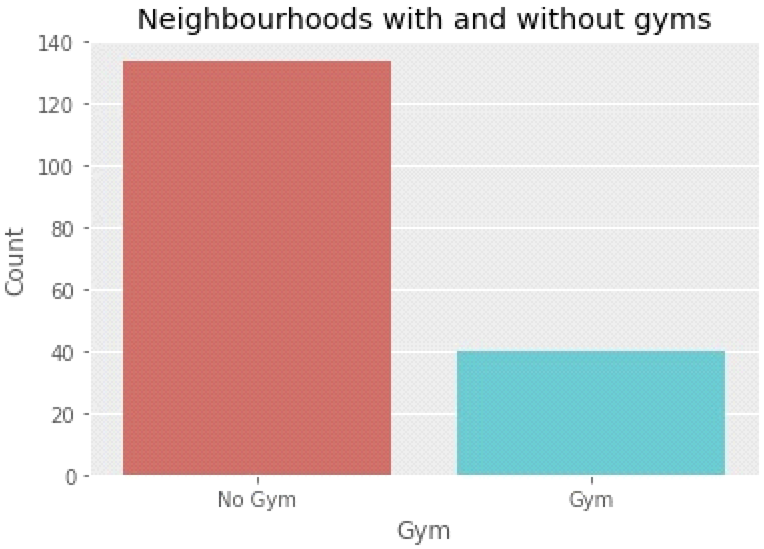
Since the variables doesn't have a linear relationship between them, the gym data is converted to one hot encoding in order to apply logistic regression.

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



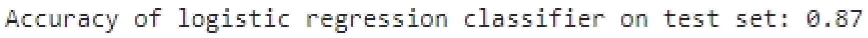
	neighbourhood	population	land_area	population_change	average_income	borough	postcode	latitude	longitude	population_score	gymcount	gymscore	gym
0	Agincourt	44577	12.45	4.6	25750	Scarborough	M1S	43.7942	-79.262	1.845247	0	0.000000	0
1	Alderwood	11656	4.94	-4.0	35239	Etobicoke	M8W	43.6024	-79.5435	0.482495	0	0.000000	0
2	Alexandra Park	4355	0.32	0.0	19687			43.6508	-79.4043	0.180273	0	0.000000	0
3	Allenby	2513	0.58	-1.0	245592			43.7114	-79.5534	0.104025	0	0.000000	0
4	Amesbury	17318	3.51	1.1	27546			43.7062	-79.4834	0.716872	1	0.806452	1

Distribution of neighborhoods with and without a gym.



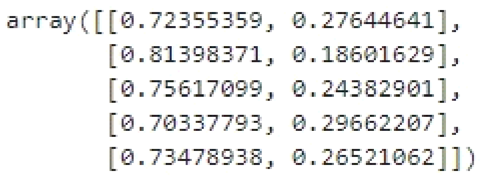
From this data, we expect to find neighborhoods that share same characteristics and features. The accuracy of the model was high and it was predicting the class 0 (No gym) with a high probability. However, the model wasn't able predict class 1 (Gym) as good as class 0.

Score of the logistic regression model.



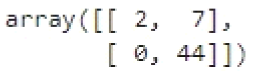
```
Accuracy of logistic regression classifier on test set: 0.87
```

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.)



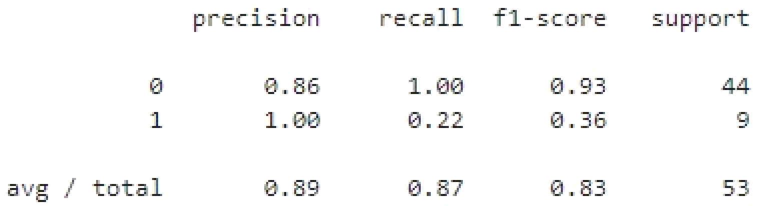
```
array([[0.72355359, 0.27644641],
       [0.81398371, 0.18601629],
       [0.75617099, 0.24382901],
       [0.70337793, 0.29662207],
       [0.73478938, 0.26521062]])
```

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.



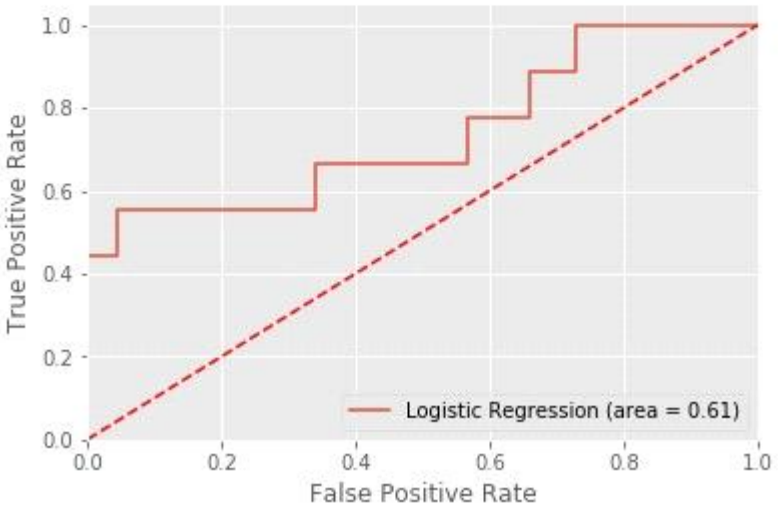
```
array([[ 2,  7],
       [ 0, 44]])
```

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



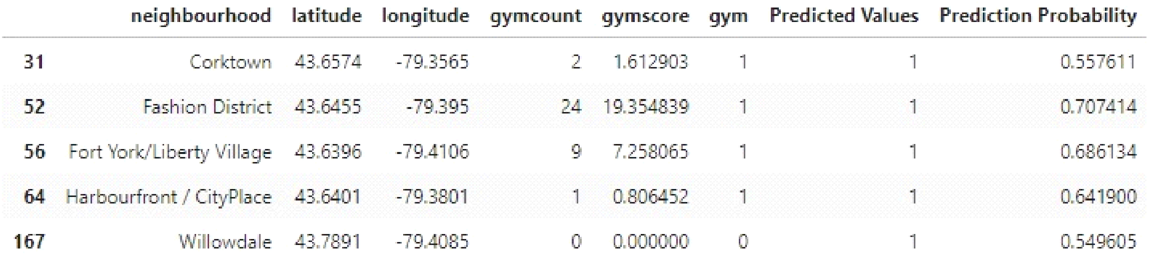
```
             precision    recall  f1-score   support

          0       0.86      1.00      0.93        44
          1       1.00      0.22      0.36         9

avg / total       0.89      0.87      0.83        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 Willowdale neighborhood predicted as having a gym, even though it doesn't have one.

Neighborhoods with a gym according to the logistic regression model.

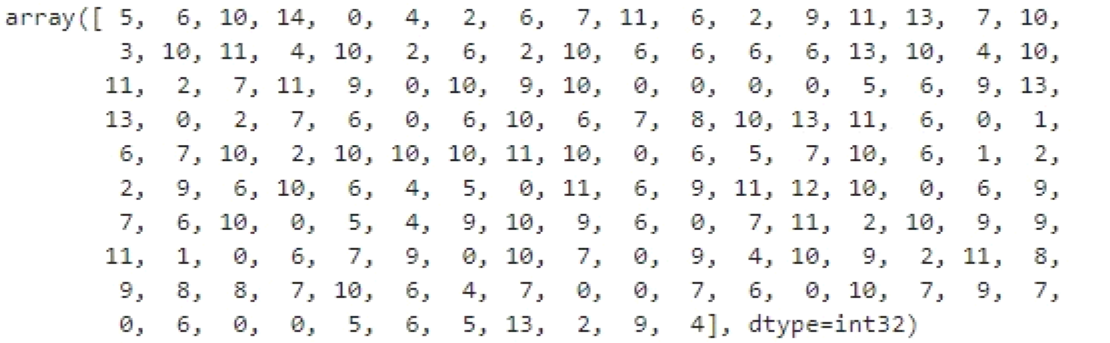


	neighbourhood	latitude	longitude	gymcount	gymscore	gym	Predicted Values	Prediction Probability
31	Corktown	43.6574	-79.3565	2	1.612903	1	1	0.557611
52	Fashion District	43.6455	-79.395	24	19.354839	1	1	0.707414
56	Fort York/Liberty Village	43.6396	-79.4106	9	7.258065	1	1	0.686134
64	Harbourfront / CityPlace	43.6401	-79.3801	1	0.806452	1	1	0.641900
167	Willowdale	43.7891	-79.4085	0	0.000000	0	1	0.549605

- **K-Means Clustering**

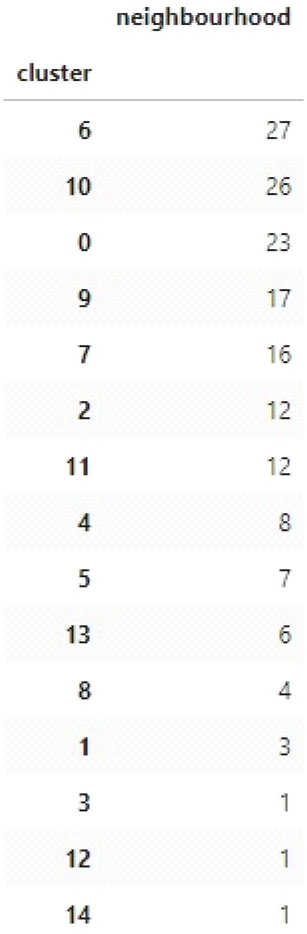
K-Means clustering will be another technique to cluster neighborhoods with shared characteristics and features. The previous neighborhood data is clustered with k=15 and fixed random_state=150 for the best results. Cluster distribution is moderately balanced and there is no bias in terms of gym count.

Cluster labels of every single neighborhood.



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

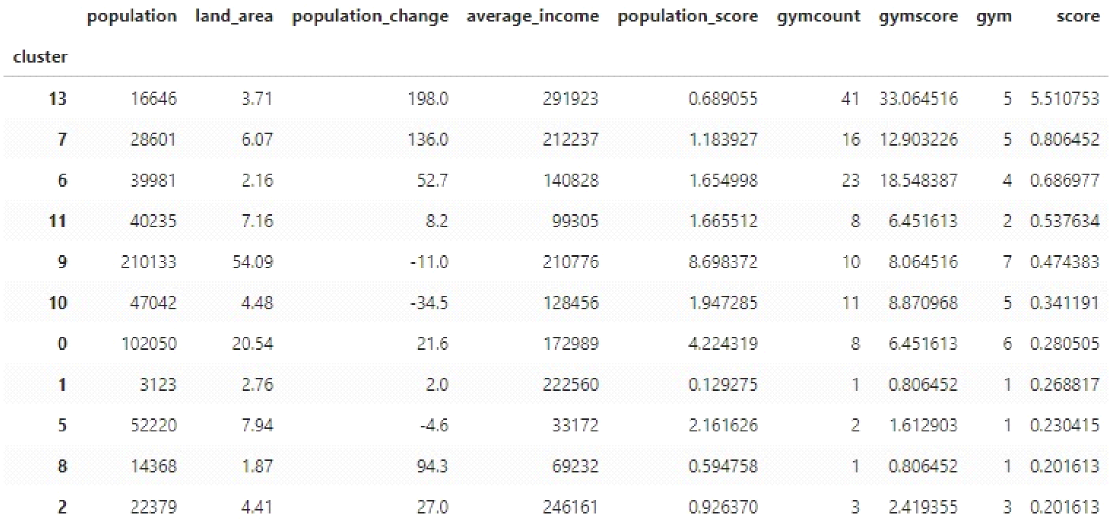
Neighborhood counts in each cluster.



	neighbourhood
cluster	
6	27
10	26
0	23
9	17
7	16
2	12
11	12
4	8
5	7
13	6
8	4
1	3
3	1
12	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.

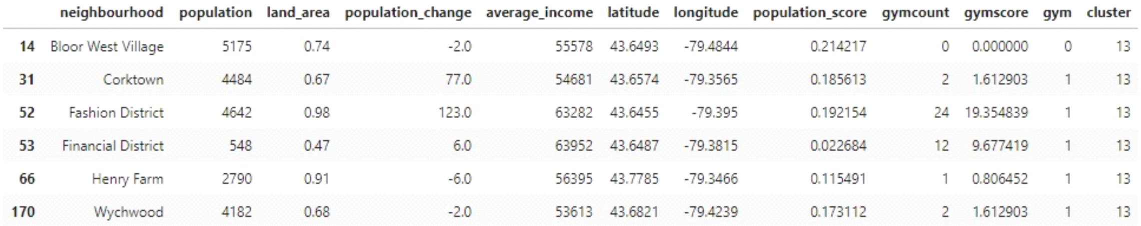
Cluster ranking by gym counts / neighborhood counts



	population	land_area	population_change	average_income	population_score	gymcount	gymscore	gym	score
cluster									
13	16646	3.71	198.0	291923	0.689055	41	33.064516	5	5.510753
7	28601	6.07	136.0	212237	1.183927	16	12.903226	5	0.806452
6	39981	2.16	52.7	140828	1.654998	23	18.548387	4	0.686977
11	40235	7.16	8.2	99305	1.665512	8	6.451613	2	0.537634
9	210133	54.09	-11.0	210776	8.698372	10	8.064516	7	0.474383
10	47042	4.48	-34.5	128456	1.947285	11	8.870968	5	0.341191
0	102050	20.54	21.6	172989	4.224319	8	6.451613	6	0.280505
1	3123	2.76	2.0	222560	0.129275	1	0.806452	1	0.268817
5	52220	7.94	-4.6	33172	2.161626	2	1.612903	1	0.230415
8	14368	1.87	94.3	69232	0.594758	1	0.806452	1	0.201613
2	22379	4.41	27.0	246161	0.926370	3	2.419355	3	0.201613

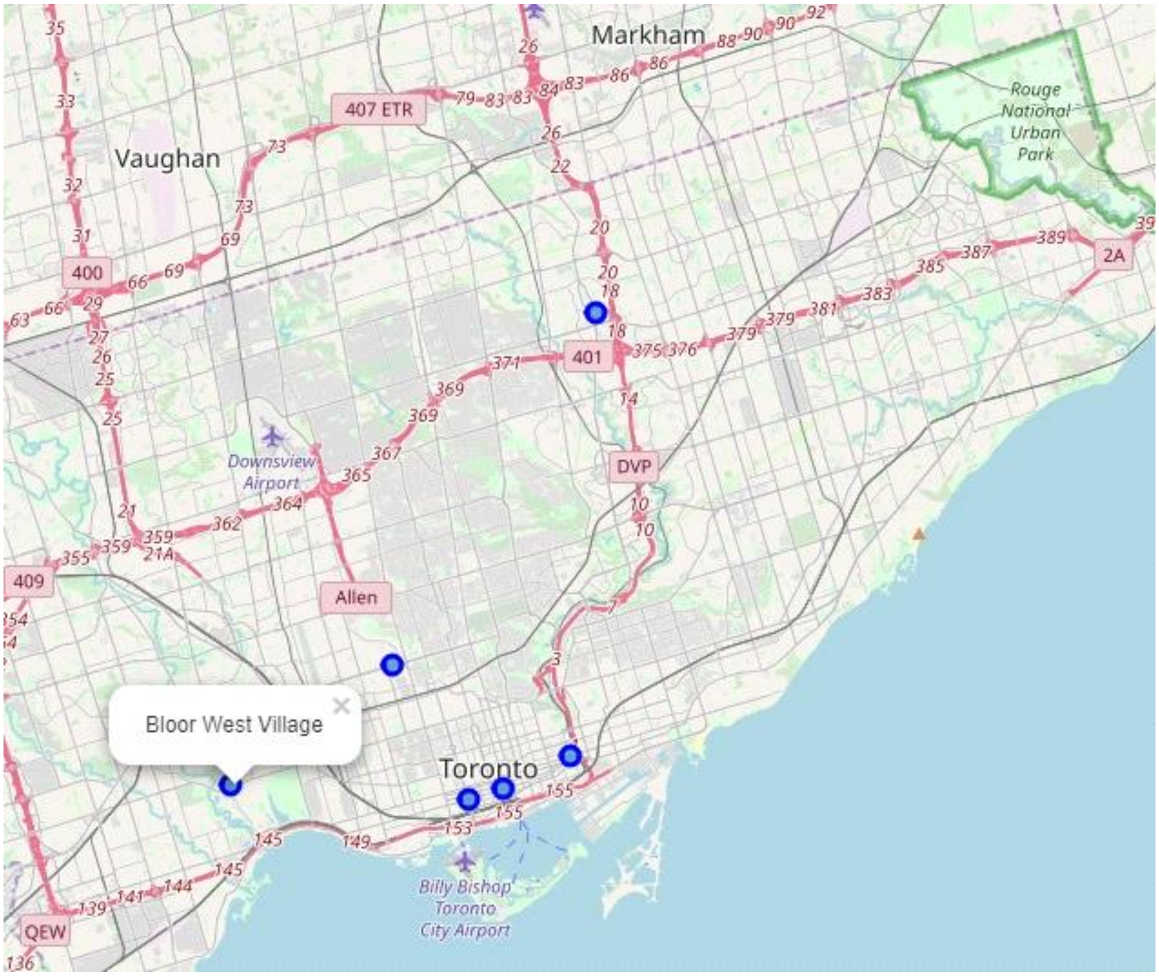
The best cluster has 6 neighborhoods. 5 of them have multiple gyms except Bloor West Village. Only Bloor West Village doesn't have a gym, but still it is clustered with those neighborhoods, so they share other features.

The best cluster in terms of gym count / neighborhood count



	neighbourhood	population	land_area	population_change	average_income	latitude	longitude	population_score	gymcount	gymscore	gym	cluster
14	Bloor West Village	5175	0.74	-2.0	55578	43.6493	-79.4844	0.214217	0	0.000000	0	13
31	Corktown	4484	0.67	77.0	54681	43.6574	-79.3565	0.185613	2	1.612903	1	13
52	Fashion District	4642	0.98	123.0	63282	43.6455	-79.395	0.192154	24	19.354839	1	13
53	Financial District	548	0.47	6.0	63952	43.6487	-79.3815	0.022684	12	9.677419	1	13
66	Henry Farm	2790	0.91	-6.0	56395	43.7785	-79.3466	0.115491	1	0.806452	1	13
170	Wychwood	4182	0.68	-2.0	53613	43.6821	-79.4239	0.173112	2	1.612903	1	13

The best cluster visualized on Toronto map

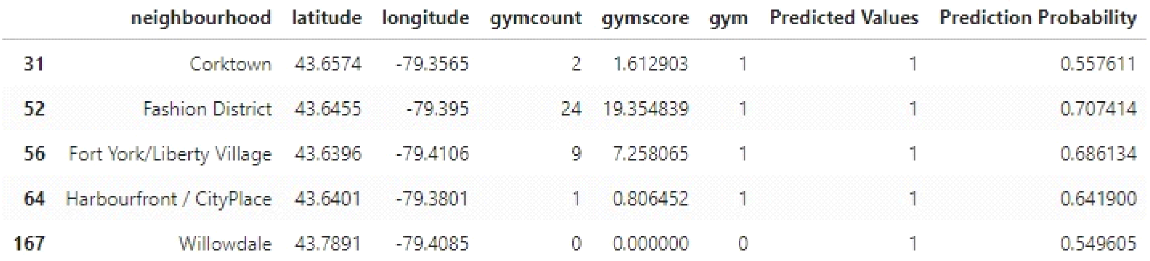


4) Result

- **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, Willowdale 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

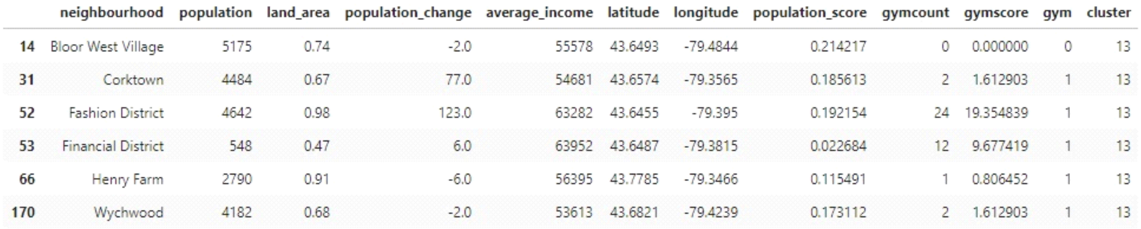


	neighbourhood	latitude	longitude	gymcount	gymscore	gym	Predicted Values	Prediction Probability
31	Corktown	43.6574	-79.3565	2	1.612903	1	1	0.557611
52	Fashion District	43.6455	-79.395	24	19.354839	1	1	0.707414
56	Fort York/Liberty Village	43.6396	-79.4106	9	7.258065	1	1	0.686134
64	Harbourfront / CityPlace	43.6401	-79.3801	1	0.806452	1	1	0.641900
167	Willowdale	43.7891	-79.4085	0	0.000000	0	1	0.549605

- **K-Means Clustering**

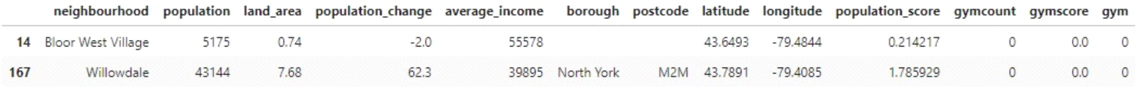
The k-means clustering model was good at clustering neighbourhoods with high number of gyms. The best cluster was selected among the clusters in terms of gym count divided by neighborhood count. The best cluster has neighborhoods with multiple gyms, but Bloor West Village doesn't have any gym. It shares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

The best cluster according to the k-means clustering model



	neighbourhood	population	land_area	population_change	average_income	latitude	longitude	population_score	gymcount	gymscore	gym	cluster
14	Bloor West Village	5175	0.74	-2.0	55578	43.6493	-79.4844	0.214217	0	0.000000	0	13
31	Corktown	4484	0.67	77.0	54681	43.6574	-79.3565	0.185613	2	1.612903	1	13
52	Fashion District	4642	0.98	123.0	63282	43.6455	-79.395	0.192154	24	19.354839	1	13
53	Financial District	548	0.47	6.0	63952	43.6487	-79.3815	0.022684	12	9.677419	1	13
66	Henry Farm	2790	0.91	-6.0	56395	43.7785	-79.3466	0.115491	1	0.806452	1	13
170	Wychwood	4182	0.68	-2.0	53613	43.6821	-79.4239	0.173112	2	1.612903	1	13

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_score	gymcount	gymscore	gym
14	Bloor West Village	5175	0.74	-2.0	55578			43.6493	-79.4844	0.214217	0	0.0	0
167	Willowdale	43144	7.68	62.3	39895	North York	M2M	43.7891	-79.4085	1.785929	0	0.0	0

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 (Willowdale) 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. Willowdale should also have a gym it is like those neighborhoods.

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.