COURSERA CAPSTONE - THE BATTLE OF NEIGHBORHOODS

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.

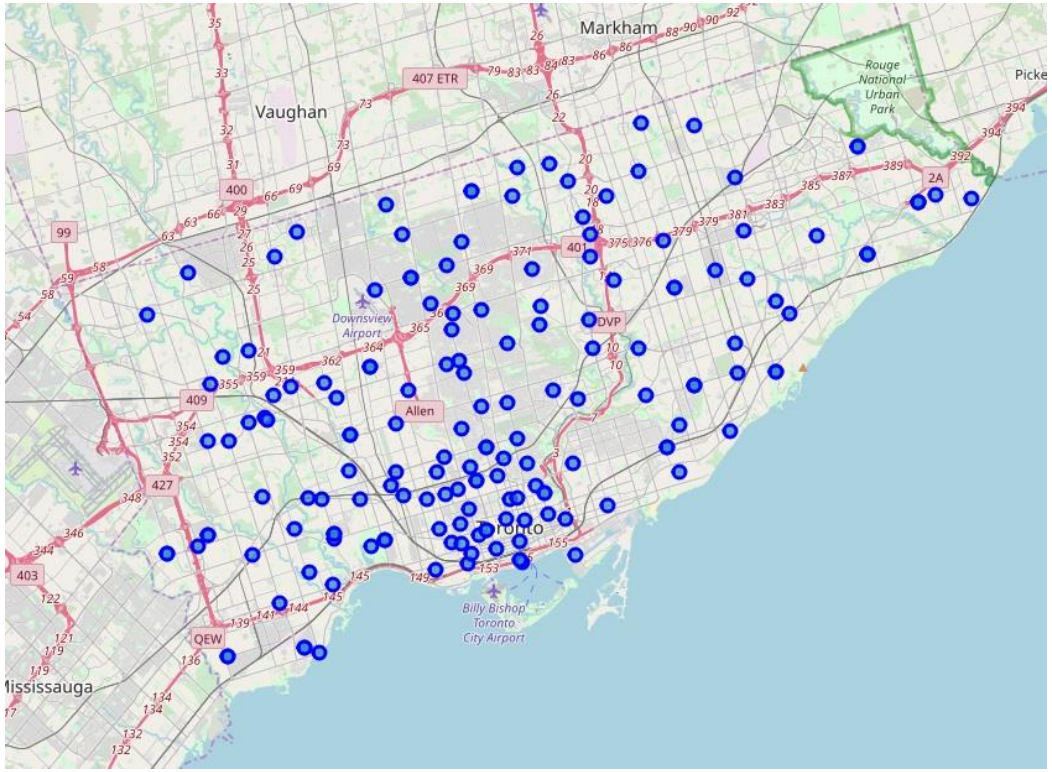
- **The business problem**: I want to open a gym in Toronto and I have to resolve the problem stated above in order to be successful. Otherwise, the business will be a failure.

- 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:** Borough and postcode of each neighborhood in Toronto
- • **Demographics of Toronto neighborhoods data set:** 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.
- • **Toronto gym data:** Gyms located in Toronto neighborhoods
- • **Geospatial Coordinates:** Latitude, longitude and postcode.

3) METHODOLOGY



- Toronto neighborhood, demographics and geospatial data merged in order to be handled easily.
- After that, the missing latitude and longitude data are found with the geopy.geocoders and inserted to table.
- Finally, all of the neighborhood data is collected, cleaned and they are ready to processed.

LOGISTIC REGRESSION

	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

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.

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.

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.

```
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)
```

- 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.
- 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.

	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

4) RESULT

- • **Logistic Regression:** The 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.

	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.

	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.