CAPSTONE PROJECT

THE BATTLE OF NEIGHBORHOODS – A COMPARATIVE STUDY OF THE EARLIER FIVE LOWER-TIER CONSTITUENT MUNICIPALITIES OF TORONTO

INTRODUCTION

Toronto is the capital of the province of Ontario and the business and financial capital of Canada. It's a multicultural city, and a growing financial hub in North America.

The City of Toronto

On January 1, 1998, Toronto was greatly enlarged, as an amalgamation of the Municipality of Metropolitan Toronto and its six lower-tier constituent municipalities; East York, Etobicoke, North York, Scarborough, York, and the original city itself. They were dissolved by an act of the Government of Ontario, and formed into a single-tier City of Toronto.

The Map of Toronto showing Old Toronto and its five lower-tier constituent municipalities

The present work was carried out with a view to suffice the understated business problem.

People, their needs and choices are highly dependent on the region they live in. The flourishment of a particular business is highly dependent on these factors. But it also depends on the market's competency. Venturing a business start-up in an area where that business has already mushroomed may not be as fruitful as implementing the same idea in a market with lesser competitors. At the same time, choosing a market without any competitor, with a view to establish monopolistic competition, without gauging the mindsets of the residents can prove to be deceptive. With all these 'words of wisdom' hovering over a businessman's head, he may wish **to get the business analytics of the different areas where he wishes to set up his new, so that he can both efficiently as well as effectively choose the most appropriate area for the same**.

The presented capstone project aims at furnishing a comparative study of the earlier five lower-tier constituent municipalities of Toronto, namely, East York, Etobicoke, North York, Scarborough and York to analyze the business prospects in each one.

DATA

It's the extraction and efficient processing of datasets, collected from various sources, that makes up the bulk of a 'data scientist'. What good is a data scientist without data? Collecting and wrangling data requires a lot of practice, patience and dedication. All datasets are unique in their own way, and each one requires a new approach.

In this capstone project, I used the data from the following sources.

1. Wikipedia: Wikipedia is a repository of a huge volume of data. The data for the different boroughs of Canada is readily available by clicking on the link given below.
   https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
   The data was extracted as a pandas dataframe, in the form as shown.

	Postal code	Borough	Neighborhood
0	M1A	Not assigned	NaN
1	M2A	Not assigned	NaN
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park / Harbourfront
5	M6A	North York	Lawrence Manor / Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park / Ontario Provincial Government
7	M8A	Not assigned	NaN
8	M9A	Etobicoke	Islington Avenue
9	M1B	Scarborough	Malvern / Rouge
10	M2B	Not assigned	NaN

We can see that many of the boroughs have not been assigned. Hence, the next thing to be done was to drop all the rows that had no boroughs assigned. Also, any neighborhood, that didn't have a value, was assigned its corresponding borough. The slash (/) for more than one neighborhood for a given borough was replaced by a comma (,). After all this, the dataframe looked like:

	Postal code	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park , Harbourfront
3	M6A	North York	Lawrence Manor , Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park , Ontario Provincial Government
5	M9A	Etobicoke	Islington Avenue
6	M1B	Scarborough	Malvern , Rouge
7	M3B	North York	Don Mills
8	M4B	East York	Parkview Hill , Woodbine Gardens
9	M5B	Downtown Toronto	Garden District, Ryerson
10	M6B	North York	Glencairn

2. https://cocl.us/Geospatial_data

   This link provided the geographical coordinates, i.e., the latitude and longitude of each Postal Code in Canada. A snapshot of the data collected from this site is given below.

	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
5	M1J	43.744734	-79.239476
6	M1K	43.727929	-79.262029
7	M1L	43.711112	-79.284577
8	M1M	43.716316	-79.239476
9	M1N	43.692657	-79.264848
10	M1P	43.757410	-79.273304

The data obtained from Wikipedia was merged with the dataset shown above, to get the following dataframe.

	Postal Code	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
5	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242
6	M1B	Scarborough	Malvern , Rouge	43.806686	-79.194353
7	M3B	North York	Don Mills	43.745906	-79.352188
8	M4B	East York	Parkview Hill , Woodbine Gardens	43.706397	-79.309937
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
10	M6B	North York	Glencairn	43.709577	-79.445073

The above dataframe was used for classifying the earlier six lower-tier constituent municipalities of Toronto.

3. Foursquare API: Foursquare is a social location service that allows users to explore the world around them. The Foursquare API allows application developers to interact with the Foursquare platform. The API itself is a RESTful set of addresses to which one can send requests, so there's really nothing to download onto the server.

Foursquare API was used in this project to get the common venues around each of the six places, by passing in the required parameters.
The dataset was of the following form:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern , Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge Hill , Port Union , Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Guildwood , Morningside , West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank
3	Guildwood , Morningside , West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store
4	Guildwood , Morningside , West Hill	43.763573	-79.188711	Big Bite Burrito	43.766299	-79.190720	Mexican Restaurant
5	Guildwood , Morningside , West Hill	43.763573	-79.188711	Enterprise Rent-A-Car	43.764076	-79.193406	Rental Car Location
6	Guildwood , Morningside , West Hill	43.763573	-79.188711	Woburn Medical Centre	43.766631	-79.192286	Medical Center
7	Guildwood , Morningside , West Hill	43.763573	-79.188711	Lawrence Ave E & Kingston Rd	43.767704	-79.189490	Intersection
8	Guildwood , Morningside , West Hill	43.763573	-79.188711	Eggsmart	43.767800	-79.190466	Breakfast Spot
9	Woburn	43.770992	-79.216917	Starbucks	43.770037	-79.221156	Coffee Shop
10	Woburn	43.770992	-79.216917	Tim Hortons	43.770827	-79.223078	Coffee Shop

METHODOLOGY

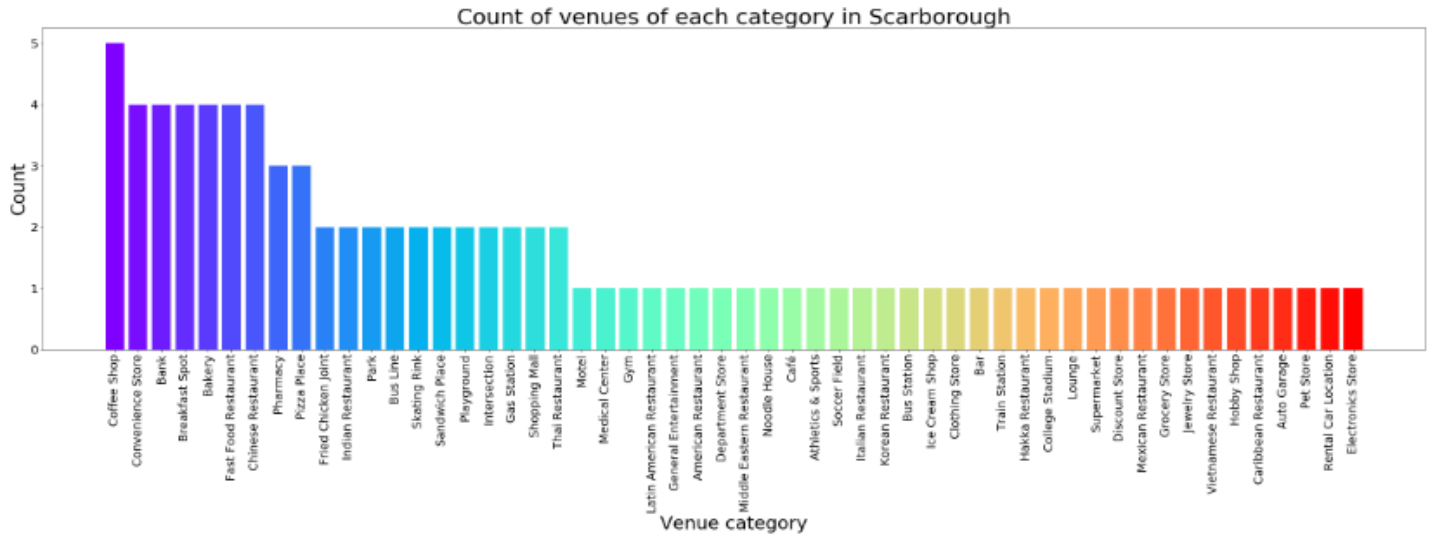
After separating the data for each borough, and fetching the venue details for each one via Foursquare API, wordclouds were made to show the popular venues in each borough. Bar graphs were plotted to help in getting the quantitative analysis of the blooming of each category of venues in the respective boroughs. On the x-axis, 'Venue Category' was taken, while on the y-axis, the count of venues belonging to the category was taken. While wordclouds would help in getting a rough idea of the popularity, bar graphs would facilitate in getting a scientific pursuit and arrive at a concrete decision.

K Nearest Neighbours was used as the Machine Learning algorithm for predicting 'Borough', given a geographical coordinate. K = 1 worked the best in this case. It could be interesting to let the machine predict the place, when it is questioned with the coordinates. The same Downtown Toronto has places with slightly differing coordinates. From a human perspective, there isn't much difference between the coordinates of different boroughs, say the coordinates of North York and East York. In such a situation, it becomes quite difficult for people to guess the boroughs, provided the coordinates are given. Seeing a machine learning algorithm perform this task, would be wonderful.

The dataframe having Postal Code, Borough, Latitude and Longitude was processed. Depending on the Latitude and Longitude, Borough was predicted. The dataset was divided into training and testing datasets by using train_test_split( ) function. 20% of the data was used for testing. Jaccard Index and F1 Score were calculated to determine the accuracy of the model.

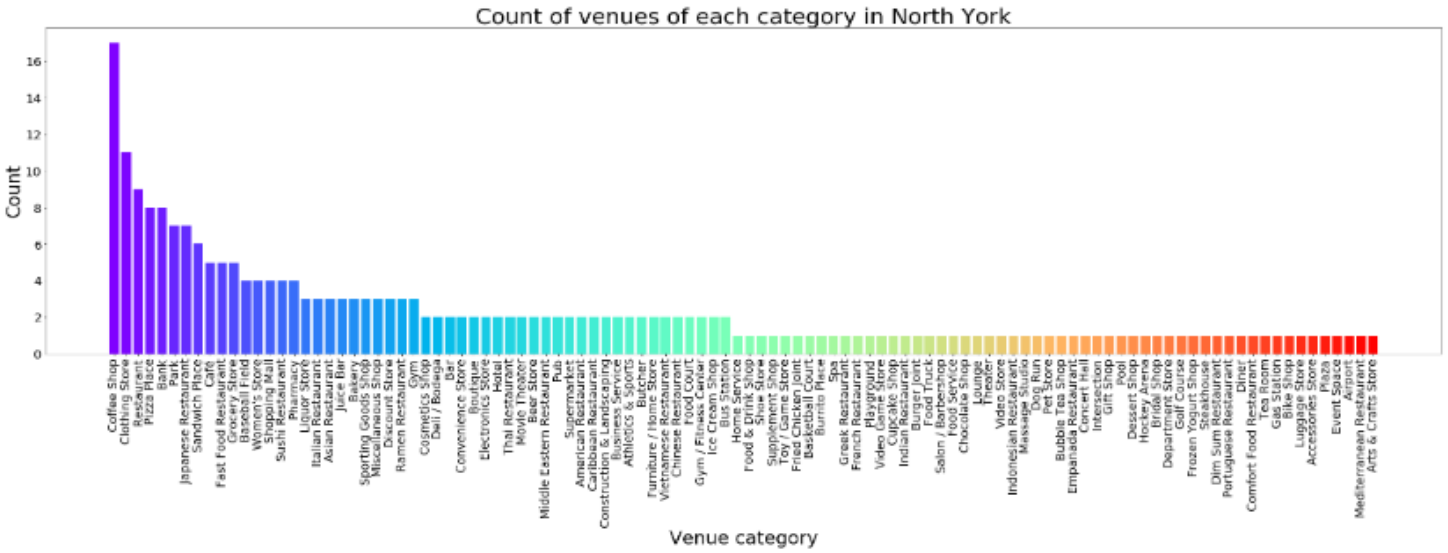
RESULTS

The bar graphs plotted showed the quantitative abundance of the different ‘venue categories’ in each borough.



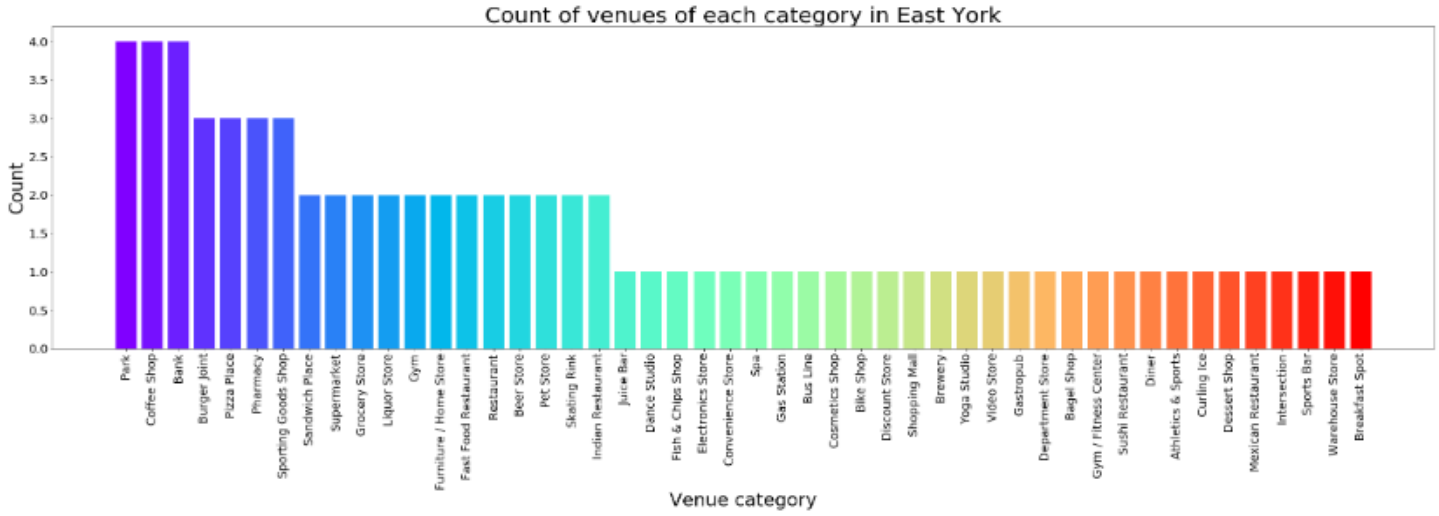
Coffee Shop is the most common venue in Scarborough.

BAR GRAPH FOR SCARBOROUGH



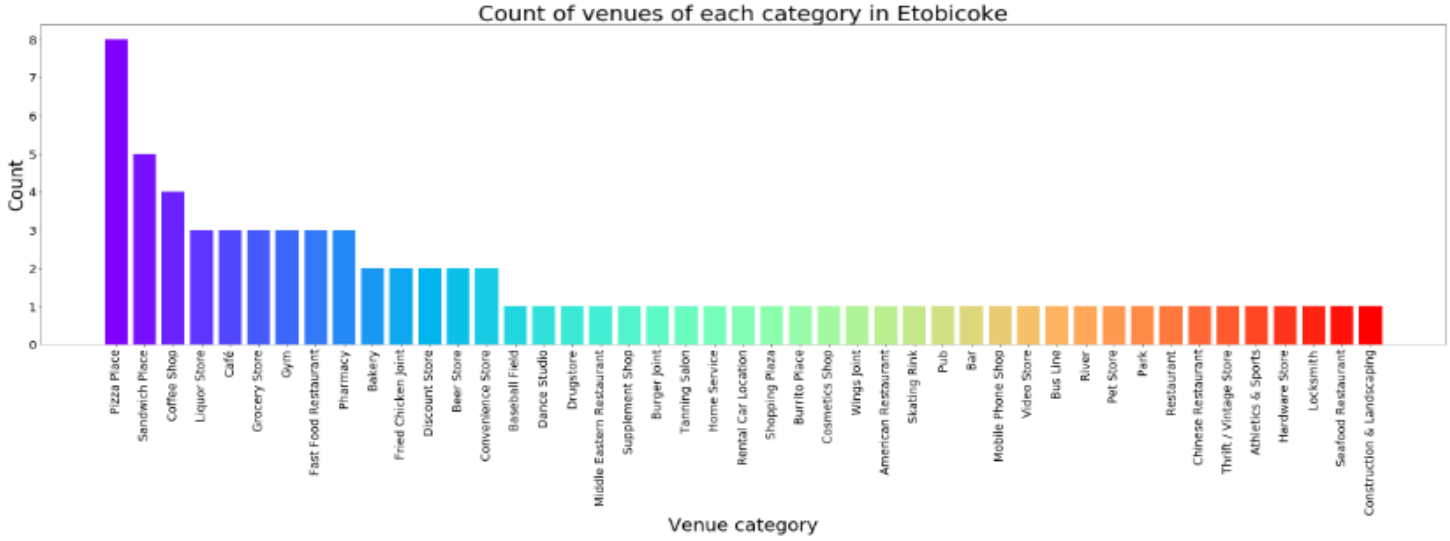
Coffee Shop is the most common venue in North York.

BAR GRAPH FOR NORTH YORK



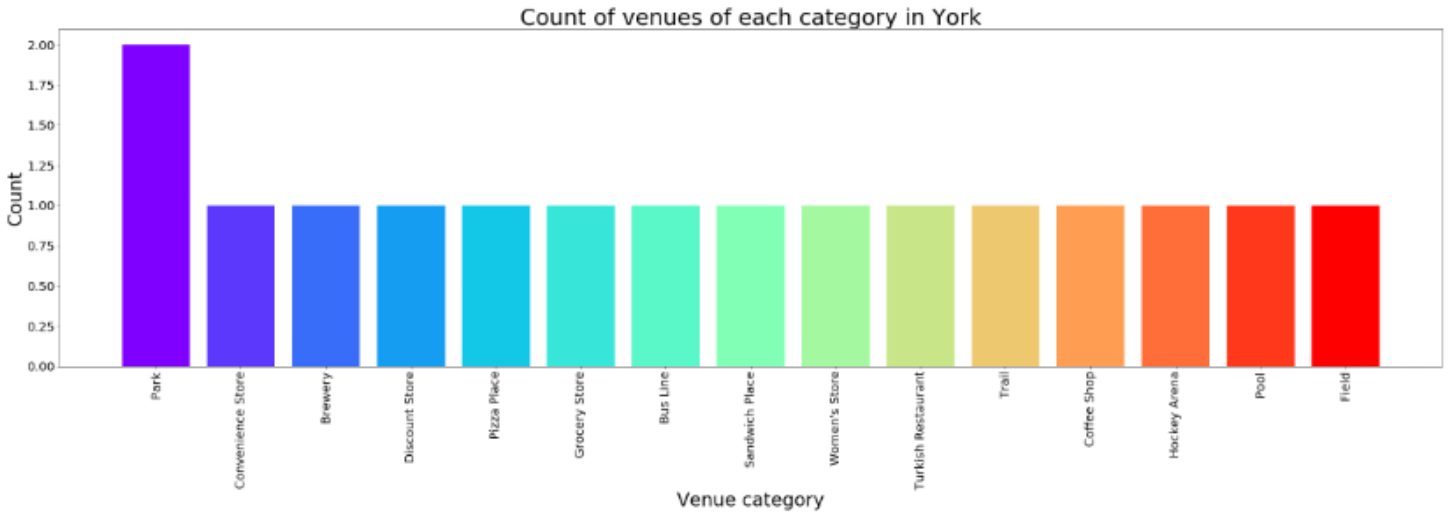
Coffee shop, bank and park form the most common venues in East York.

BAR GRAPH FOR EAST YORK



Pizza Place is the most common venue in Etobicoke.

BAR GRAPH FOR ETOBICOKE



Last but not the least, parks are the most common venue at York.

BAR GRAPH FOR YORK

MACHINE LEARNING MODEL

We can observe that the accuracy of the model on the training data set is 100% while on the test data set is 85.7%.

In [137]:
```
k = 1
neigh = KNeighborsClassifier(n_neighbors = k).fit(X_train,y_train)
yhat = neigh.predict(X_test)
print("Train set Accuracy: ", metrics.accuracy_score(y_train, neigh.predict(X_train)))
print("Test set Accuracy: ", metrics.accuracy_score(y_test, yhat))
```

```
Train set Accuracy:  1.0
Test set Accuracy:  0.8571428571428571
```

Comparing few rows of the predicted data set and original test data set:

In [141]:
```
print(yhat[0:10])
```

```
['East Toronto' 'Etobicoke' 'Downtown Toronto' 'Downtown Toronto'
 'Downtown Toronto' 'North York' 'North York' 'Scarborough' 'East Toronto'
 'Central Toronto']
```

In [142]:
```
print(y_test[0:10])
```

```
['East York' 'Etobicoke' 'Downtown Toronto' 'Downtown Toronto'
 'Downtown Toronto' 'North York' 'North York' 'Scarborough' 'East Toronto'
 'Central Toronto']
```

Values for:

- Jaccard Similarity Index : 0.857
- F1 Score : 0.849

In [140]:
```
yhat = neigh.predict(X_test)
jss_knn = jaccard_similarity_score(y_test , yhat)
f1_knn = f1_score(y_test, yhat, average = 'weighted')
print("Jaccard Similarity Index for K-Nearest Neighbors : ",jss_knn)
print("F1 Score for K-Nearest Neighbors : ",f1_knn)
```

```
Jaccard Similarity Index for K-Nearest Neighbors :  0.8571428571428571
F1 Score for K-Nearest Neighbors :  0.8492063492063492
```

DISCUSSIONS

It can be seen that York is not much developed with regards to business avenues, while North York is the most developed one in that case. In many of the boroughs, coffee shops topped the list by being the most common venue category. Our KNN Machine Learning algorithm worked pretty well in predicting the borough.

In future, work may be done to further classify the venue categories in different areas of a borough. The borough data can also be linked with the income data to reveal its socio-economic aspects.

CONCLUSION

The present work reveals the business prospects in the five lower-tier constituent municipalities of Toronto, namely East York, Etobicoke, North York, Scarborough, York. This could be used to further develop these places, by opening new avenues for the ‘venue categories’ that are lesser in number. It can help in the overall development of these areas. The machine learning addition in this project can help in knowing the place, given its coordinates, which is not a human’s cup of tea.