The Battle of Neighborhoods

Title: Business Venue Recommender System in Toronto

IBM CAPSTONE PROJECT

Problem Background :

Toronto is the capital of the province of Ontario, Canada. It is a major Canadian city along Lake Ontario's northwestern shore. It's a dynamic metropolis with a core of soaring skyscrapers. It is also the fastest growing city in North America. Toronto is an international centre of business, finance, arts, and culture, and is recognized as of the most multicultural and cosmopolitan cities of the world.

Problem Description :

A businessman already has a restaurant being operated successfully in one neighborhood of Ontario. Suppose, he/she wants to increase their revenue by opening another branch of the restaurant in other part of the city. In such situation, the type of neighborhood plays an important in choosing an optimum location for the new branch. Factors like the kinds of venues in the neighborhood, population of the neighborhood, income of the neighborhood and so on have a significant effect on the location chosen for the new branch.

Our aim would be to find a location similar to the location of the original branch to minimize the risks.

Target Audience :

Target Audience for this project is not only limited to business men with restaurant businesses but also other businesses like Construction, bookstore etc. This project can be used by anyone who is looking up to expand their business in neighborhoods with some similar characteristics.

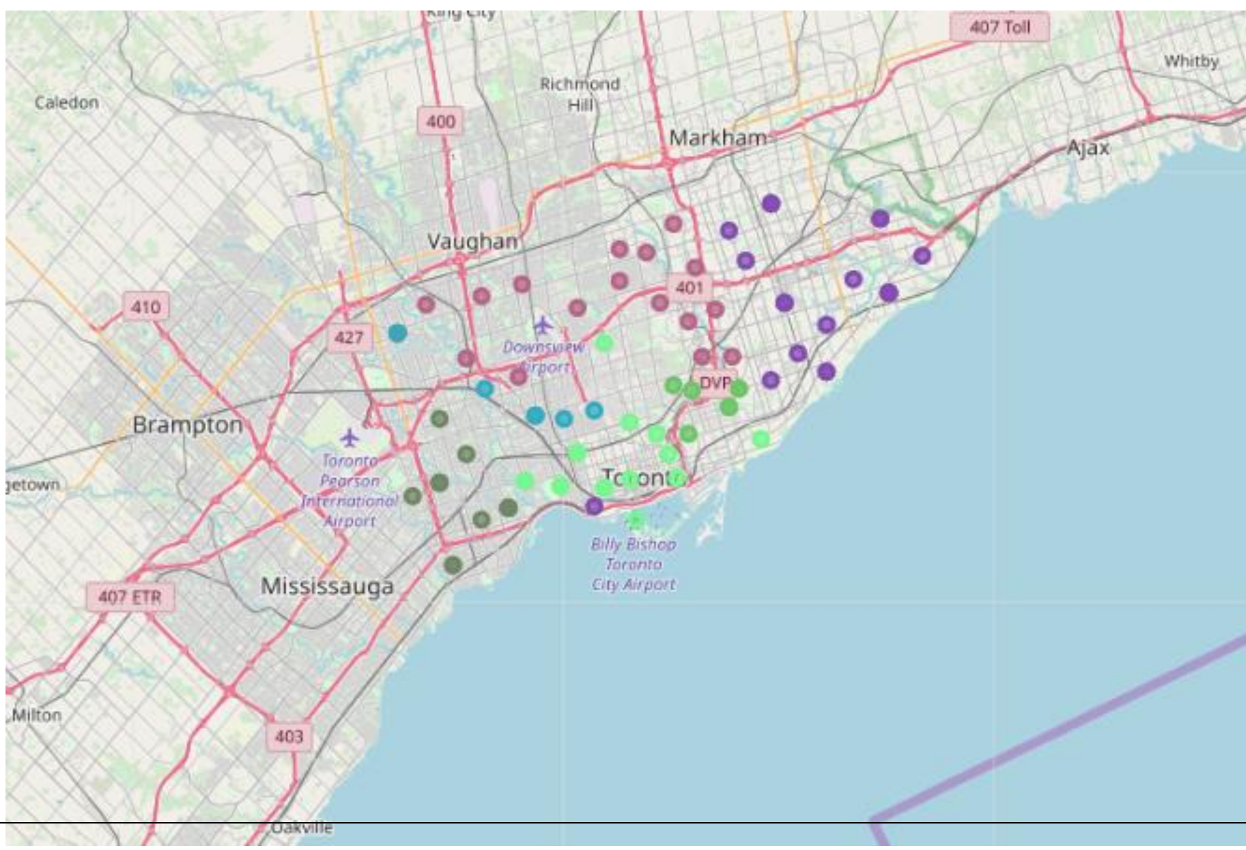
Data Requirements :

For a Recommender system, we need data and lots of data. Data can answer question which are unimaginable and non answerable by humans because humans do not have the tendency to analyze such large dataset and produce analytics to find solutions.

1. We will need information about all the neighborhoods and the boroughs of the city of Toronto. We would also need each neighborhood's latitude and longitude information. We would also need other information like income, and population of each neighborhood. I found the neighborhood, income, and population information from here:
   https://www.toronto.ca/ext/open_data/catalog/data_set_files/2016_neighbourhood_profiles.csv
   I mapped each neighborhood to its postal code and borough from the Wikipedia page of Toronto City.
   I found the latitude and longitude from here mapped to each postal code:
   http://cocl.us/Geospatial_data

	Post Code	Borough	Neighbourhood	Population	Income
0	M4K	East York	Broadview North	11499	44557
1	M4C	East York	Danforth East York	17180	51846
2	M4G	East York	Bennington	16828	125564
3	M4B	East York	O'Connor-Parkview	18675	43907
4	M4H	East York	Thorncliffe Park	21108	28875

	Post Code	Borough	Neighbourhood	Population	Income	Latitude	Longitude
0	M4K	East York	Broadview North	11499	44557	43.679557	-79.352188
1	M4C	East York	Danforth East York	17180	51846	43.695344	-79.318389
2	M4C	East York	Woodbine-Lumsden	7865	47710	43.695344	-79.318389
3	M4G	East York	Bennington	16828	125564	43.709060	-79.363452
4	M4B	East York	O'Connor-Parkview	18675	43907	43.706397	-79.309937
5	M4B	East York	Woodbine Corridor	12541	55199	43.706397	-79.309937
6	M4H	East York	Thorncliffe Park	21108	28875	43.705369	-79.349372
7	M8W	Etobicoke	Alderwood	12054	47709	43.602414	-79.543484
8	M8W	Etobicoke	Long Branch	10084	47384	43.602414	-79.543484
9	M8Y	Etobicoke	Edenbridge-Humber Valley	15535	101551	43.636258	-79.498509
10	M8Y	Etobicoke	New Toronto	11463	44101	43.636258	-79.498509
11	M8Y	Etobicoke	Queensway	25051	64140	43.636258	-79.498509
12	M9B	Etobicoke	West Deane	18588	47002	43.650943	-79.554724
13	M9B	Etobicoke	Martingrove	22156	44177	43.650943	-79.554724



2. We would also need information of venues, their longitude and location of each venue. This is where FourSquare API comes into play. Use of foursquare is focused to fetch nearest venue locations so that we can use them to form a cluster. Foursquare API leverages the power of finding nearest venues in a radius (in my case : 500mts) and also corresponding coordinates, venue location and names.

	Neighborhood	Borough	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Broadview North	East York	43.679557	-79.352188	Pantheon	43.677621	-79.351434	Greek Restaurant
1	Broadview North	East York	43.679557	-79.352188	Dolce Gelato	43.677773	-79.351187	Ice Cream Shop
2	Broadview North	East York	43.679557	-79.352188	MenEssentials	43.677820	-79.351265	Cosmetics Shop
3	Broadview North	East York	43.679557	-79.352188	Cafe Fiorentina	43.677743	-79.350115	Italian Restaurant
4	Broadview North	East York	43.679557	-79.352188	La Diperie	43.677530	-79.352295	Ice Cream Shop

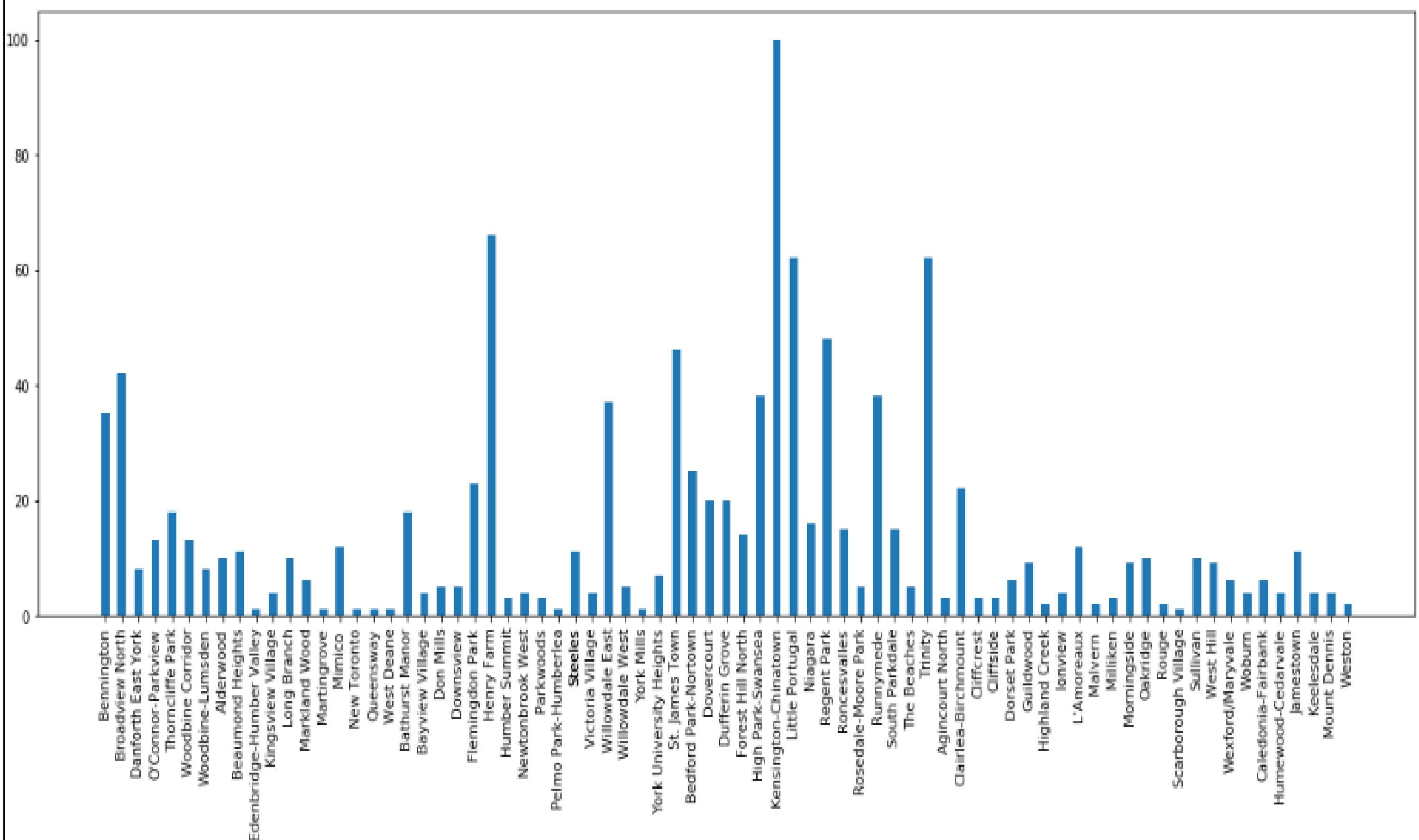
3.Methodology :

Exploratory Analysis :

Scrapping the data from different sources and then combining it to form a single-ton dataset is a difficult task. To do so, we need to explore the current state of dataset and then list up all the features needed to be fetched.

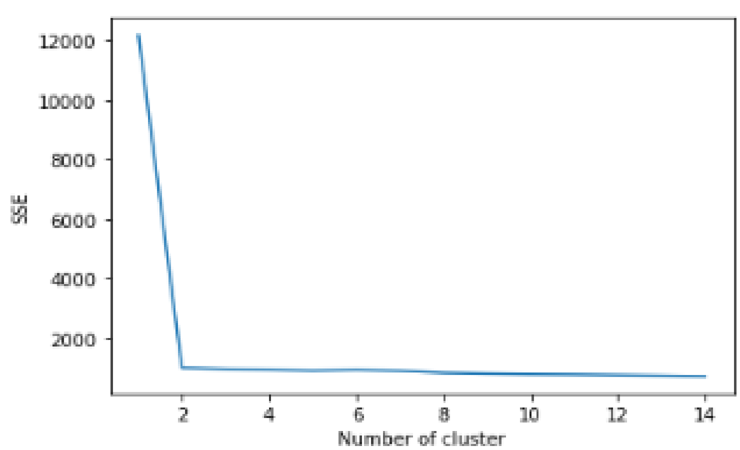
Exploring dataset is important because it gives you initial insights and may help you get a partial data of the answers that you are looking to find from the data.

While exploring the dataset I found out that Kensingtown - Chinatown has the most number of venues and Henry Farm has the second most number of venues.



Inferential Analysis :

After some initial exploratory analysis, the plan was to cluster neighborhoods which have similar features. For this, K-means clustering would come in use. K-means clustering is a type of unsupervised learning which is used when you have unlabeled data. The algorithm works iteratively to assign each data point to one K groups based on the features that are provided. Now the question arises, how many clusters should there be? For that, we used elbow graph to find the optimal value of K.



As evident front the graph, we get an optimal value of K as 2. So we fit our dataset and get two clusters which have similar feature set. We find the cluster number of the neighborhood where the business man already has a branch open i.e. we want to find a neighborhood similar to that one. All the neighborhoods which have a similar cluster number are our potential neighborhoods

To recommend neighborhoods, we need to factor in the income and population of the neighborhoods as well. So we merge the population and income datasets to our main dataset and create a ranking (Normalized Population*0.5+Normalized Income*0.35+Number of Non Coffee Shops*0.1). We create a range of ranking close to the ranking of our sample neighborhood. All the neighborhoods within that range are our recommended neighborhoods.

Result :

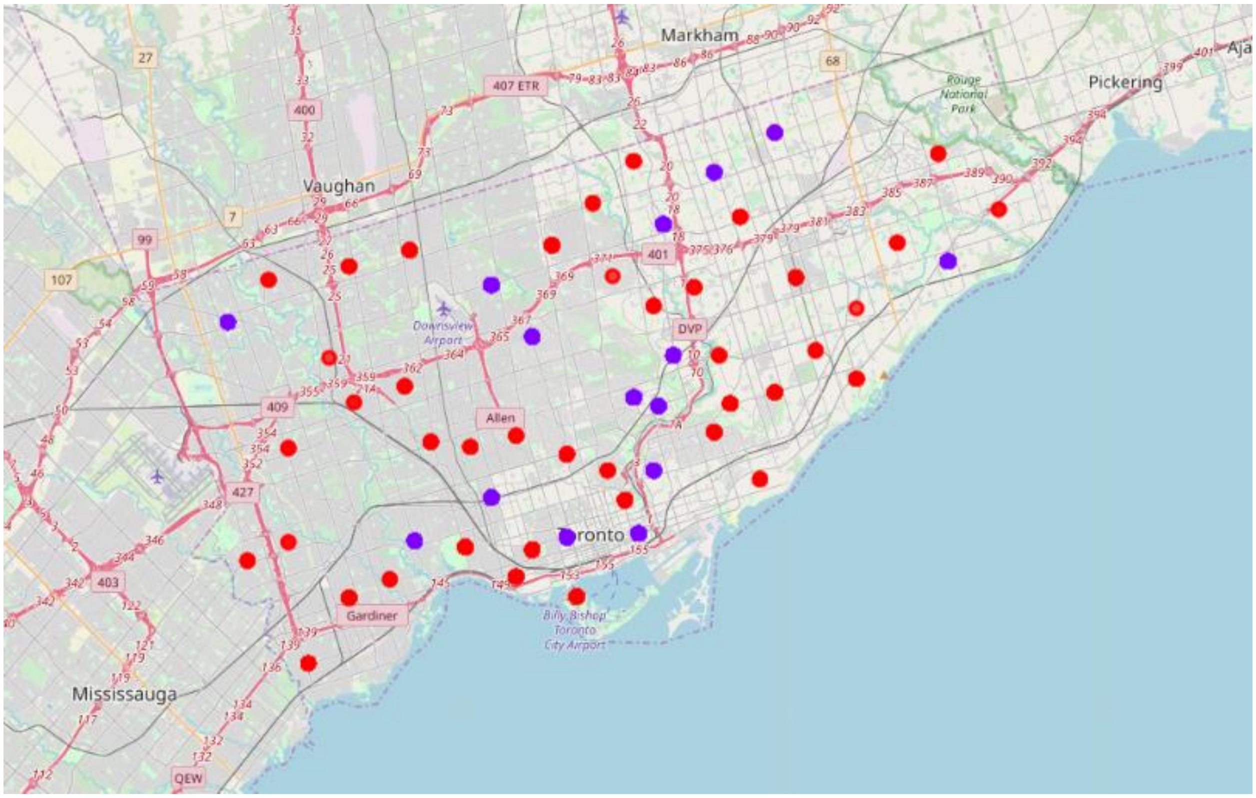
The result of the recommender system is that it produces a list of neighborhoods with their most common type of venues. During the runtime of model, a simulation was done by taking 'Regent Park' as the neighborhood and then processed through our model so that it would recommend neighborhoods with similar characters that of 'Regent Park'.

The following image shows the result.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Rank
0	St. James Town	Coffee Shop	Restaurant	Italian Restaurant	0.230689
1	Bathurst Manor	Coffee Shop	Pizza Place	Middle Eastern Restaurant	0.275586
3	Beaumond Heights	Grocery Store	Pharmacy	Fast Food Restaurant	0.221638

Discussion :

Similar neighborhoods must be dumped in the right cluster. The following graph shows the clusters on the map.



Choosing number of clusters is an important task. Diverse results are produced with different clusters. Some may be overfitted and some may be underfitted. Hence analysis of cluster must be done. Refer to the elbow graph in the methodology section.

Conclusion :

The recommender system is a system that considers factors such as population, income and use of Foursquare API to determine near by venues. It is a powerful data driven model whose accuracy will increase with more data. It finds a neighborhood with similar data features so that the business is as successful as the original venue and the risks are relatively reduced to open up a new branch.