Capstone Project Report

The Battle of Neighborhoods

Soumyadeep Bhattacharjee

Contents

1. Introduction

1.1 Background

In order to find the perfect house, one needs to find the perfect neighborhood. This is no small feat, especially if a person is about to move into a completely unknown area and cannot discern one neighborhood from another.

Even for a person who wishes to open a store or a Restaurant, he needs to understand the vibe of the neighborhood. An area with less footfall might not be a profitable zone to start a business. Similarly, regions which have limited Restaurants or stores could be a prospective area to open a new Restaurant, as the business might prosper owing to lesser market competition.

1.2 Problem Statement

With the ever-increasing demand for real estate, the decision to choose a neighborhood to build a House or start a Restaurant, needs to be made quickly and accurately.

To allow for an easier decision-making process, neighborhoods can be clustered into prospective zones based on popularity/availability of venues. These clusters would help in identifying the ideal neighborhood based on the need.

To better illustrate this, a sample segmentation has been performed on the neighborhoods in Toronto leveraging the Foursquare API and all other relevant data collected throughout the course of this project.

1.3 Interest Groups

- Expats wishing to relocate to a new area
- Budding entrepreneurs or Businesspersons wishing to find a neighborhood to start a new Store, Mall or Restaurant
- Real Estate agencies who wish to identify zones and ascertain House prices based on the popularity of the neighborhood.

2. Acquiring and Pre-processing data

2.1 Data Acquisition

Data regarding the List of postal codes, Boroughs and Neighborhoods of Toronto, Canada was obtained by scraping data from the following Wikipedia page:
[List of postal codes of Canada: M]

The obtained response was transformed into a pandas dataframe:

	PostalCode	Borough	Neighborhood
0	M1A\n	Not assigned\n	\n
1	M2A\n	Not assigned\n	\n
2	M3A\n	North York\n	Parkwoods\n
3	M4A\n	North York\n	Victoria Village\n
4	M5A\n	Downtown Toronto\n	Regent Park / Harbourfront\n
...	...	...	...
175	M5Z\n	Not assigned\n	\n
176	M6Z\n	Not assigned\n	\n
177	M7Z\n	Not assigned\n	\n
178	M8Z\n	Etobicoke\n	Mimico NW / The Queensway West / South of Bloo...
179	M9Z\n	Not assigned\n	\n

	PostalCode	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

Geospatial Data corresponding to each postal code was obtained from the link provided in the course: [Geospatial_data]

All nearby venues within 1 KM radius of each neighborhood was collected using the explore call to the Foursquare API:

	Neighborhood	Venue	Latitude	Longitude	Category
0	Parkwoods	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant
1	Parkwoods	Brookbanks Park	43.751976	-79.332140	Park
2	Parkwoods	Tim Hortons	43.760668	-79.326368	Café
3	Parkwoods	A&W	43.760643	-79.326865	Fast Food Restaurant
4	Parkwoods	Bruno's valu-mart	43.746143	-79.324630	Grocery Store
5	Parkwoods	High Street Fish & Chips	43.745260	-79.324949	Fish & Chips Shop
6	Parkwoods	Food Basics	43.760549	-79.326045	Supermarket
7	Parkwoods	Shoppers Drug Mart	43.745315	-79.325800	Pharmacy
8	Parkwoods	Shoppers Drug Mart	43.760857	-79.324961	Pharmacy
9	Parkwoods	Variety Store	43.751974	-79.333114	Food & Drink Shop
10	Parkwoods	Pizza Pizza	43.760231	-79.325666	Pizza Place
11	Parkwoods	DVP at York Mills	43.758899	-79.334099	Road
12	Parkwoods	TTC Stop #09083	43.759655	-79.332223	Bus Stop
13	Parkwoods	TTC Stop 9083	43.759251	-79.334000	Bus Stop
14	Parkwoods	Sandover Park	43.760277	-79.333305	Park
15	Parkwoods	TTC Stop #9075	43.757596	-79.338155	Train Station
16	Parkwoods	Dollarama	43.760341	-79.325519	Discount Store
17	Parkwoods	Parkwoods Coin Laundry	43.760386	-79.324894	Laundry Service
18	Parkwoods	Spicy Chicken House	43.760639	-79.325671	Chinese Restaurant
19	Parkwoods	La Notre	43.760704	-79.325396	Coffee Shop
20	Parkwoods	Underhill Mini Mart Convenience	43.745836	-79.324835	Convenience Store
21	Parkwoods	Parkwoods Village Centre	43.760735	-79.324873	Shopping Mall
22	Parkwoods	Family Food Fair Convenience	43.760620	-79.324459	Convenience Store
23	Parkwoods	Broadlands Skating Rink	43.746689	-79.322678	Skating Rink
24	Parkwoods	Parkway Valley Tennis Club	43.754481	-79.318285	Tennis Court

2.2 Data Cleaning

The data obtained by scraping the Wikipedia page was not ready to be used for analysis straightaway. The following pre-processing operations were performed on data before it could be used for further analysis:

- Removing newline characters
- Replace '/' with a ',' as a separator for multiple neighborhoods
- Removing Rows where the Borough is 'Not assigned'
- Removing extra spaces from the 'PostalCode' column
- Merging the two dataframes based on the PostalCode Column

	PostalCode	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
...	...	...	...	...	...
98	M8X	Etobicoke	The Kingsway , Montgomery Road , Old Mill North	43.653654	-79.506944
99	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160
100	M7Y	East Toronto	Business reply mail Processing CentrE	43.662744	-79.321558
101	M8Y	Etobicoke	Old Mill South , King's Mill Park , Sunnylea ,...	43.636258	-79.498509
102	M8Z	Etobicoke	Mimico NW , The Queensway West , South of Bloo...	43.628841	-79.520999

Similarly, the venues data returned from the Foursquare API, was also cleaned removing any unnecessary spaces and characters that would hinder further comparison and analysis.

3. Methodology

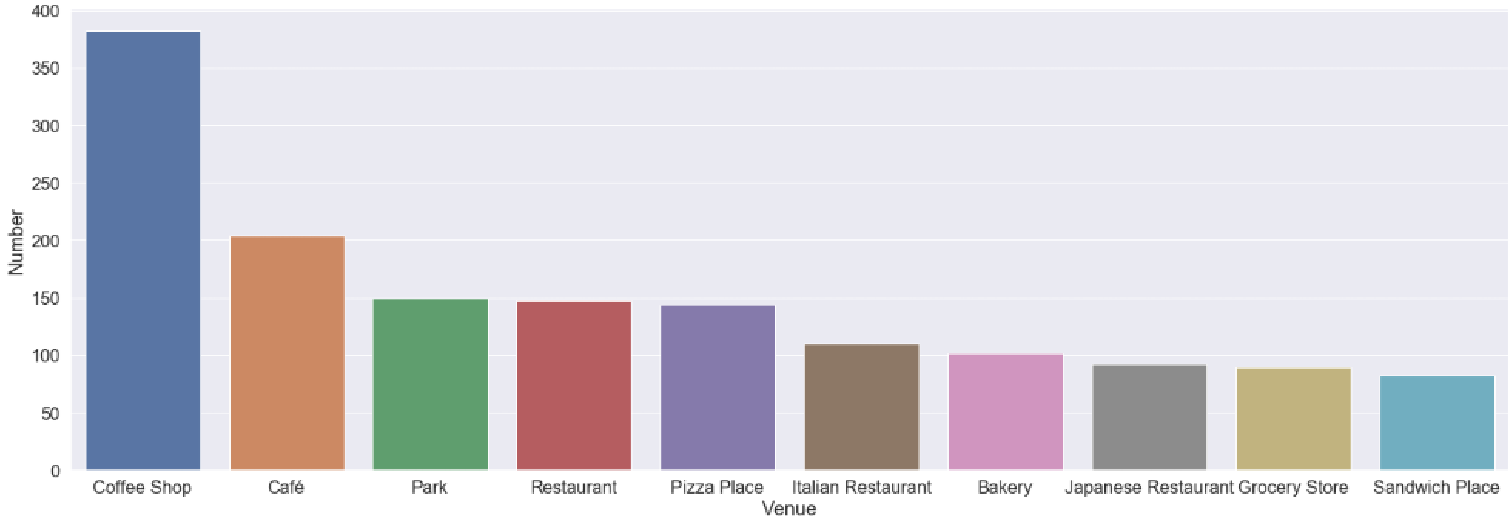
3.1 Exploratory Data Analysis

In order to identify the most frequent venues to choose to cluster neighborhoods, exploratory data analysis was performed on the list of venues obtained from the Foursquare API query.

- All unique venues were identified, and their frequency of occurrence was stored in a dataframe.

	Venue	Frequency
0	Coffee Shop	382
1	Café	204
2	Park	150
3	Restaurant	148
4	Pizza Place	144
5	Italian Restaurant	110
6	Bakery	102
7	Japanese Restaurant	92
8	Grocery Store	90
9	Sandwich Place	82
10	Bar	80
11	Bank	76
12	Gym	74

- The top 10 most frequent locations were plotted as a bar graph to offer a better visual understanding



- We can note that most of the top venues are eateries. Hence, additional categories like ‘Gym’, ‘Grocery Store’, ‘Bank’ and ‘Pharmacy’ were included as features to create a more inclusive feature set. A dataframe was constructed which showed the number of occurrences of each venue per neighborhood.

	Neighborhood	Restaurant	Park	Bank	Pharmacy	Coffee	Café	Bar	Pizza	Gym	Grocery Store	Total
0	Parkwoods	3	3	0	2	1	1	0	1	0	1	12
1	Victoria Village	1	1	0	0	2	0	0	0	2	0	6
2	Regent Park , Harbourfront	20	4	1	1	15	4	1	1	2	1	50
3	Lawrence Manor , Lawrence Heights	13	1	1	0	3	0	0	0	2	1	21
4	Queen's Park , Ontario Provincial Government	24	4	0	0	8	2	6	2	1	1	48
5	Islington Avenue	0	1	1	2	0	1	0	0	0	1	6
6	Malvern , Rouge	5	0	1	0	2	0	0	0	0	0	8
7	Don Mills	24	1	3	0	7	1	1	2	4	0	43
8	Parkview Hill , Woodbine Gardens	2	0	1	1	1	0	0	2	1	0	8
9	Garden District, Ryerson	27	2	0	0	10	2	5	1	2	1	50

3.2 Clustering the data using K-Means Algorithm

To cluster similar neighborhoods based on the venues/amenities available around a 1 Kilometer radius, we will be clustering similar neighborhoods using K - means clustering algorithm – an unsupervised machine learning algorithm that groups data based on a predefined cluster size.

For the present use case, we will use a cluster size of 3. Each of the clusters would signify the popularity level of a neighborhood:

- Low
- Medium
- High

Upon clustering, the cluster labels along with the Co-ordinates and Total Venues for each neighborhood were stored in a dataframe to allow visualizing the clusters on the Map of Toronto.

Neighborhood	ClusterLabel	Latitude	Longitude	Total
Upper Rouge	0	43.836125	-79.205636	0
Northwest	0	43.706748	-79.594054	1
Humberlea , Emery	0	43.724766	-79.532242	1
Rouge Hill , Port Union , Highland Creek	0	43.784535	-79.160497	2
York Mills , Silver Hills	0	43.757490	-79.374714	3
Old Mill South , King's Mill Park , Sunnylea ,...	0	43.636258	-79.498509	5
CN Tower , King and Spadina , Railway Lands , ...	0	43.628947	-79.394420	5
Cliffside , Cliffcrest , Scarborough Village West	0	43.716316	-79.239476	6
Humber Summit	0	43.756303	-79.565963	6
Del Ray , Mount Dennis , Keelsdale and Silvert...	0	43.691116	-79.476013	6
Birch Cliff , Cliffside West	0	43.692657	-79.264848	6

4. Results

Running the K-means clustering algorithm on the Neighborhood data allows us to view each cluster formed. to see which neighborhoods was assigned to each of the 3 clusters.

Looking into the neighborhoods in Cluster 0, we can easily identify that the 'Total' availability of venues is **Low**:

Neighborhood	ClusterLabel	Latitude	Longitude	Total
Upper Rouge	0	43.836125	-79.205636	0
Northwest	0	43.706748	-79.594054	1
Humberlea , Emery	0	43.724766	-79.532242	1
Rouge Hill , Port Union , Highland Creek	0	43.784535	-79.160497	2
York Mills , Silver Hills	0	43.757490	-79.374714	3
Old Mill South , King's Mill Park , Sunnylea ,...	0	43.636258	-79.498509	5
CN Tower , King and Spadina , Railway Lands , ...	0	43.628947	-79.394420	5
Cliffside , Cliffcrest , Scarborough Village West	0	43.716316	-79.239476	6

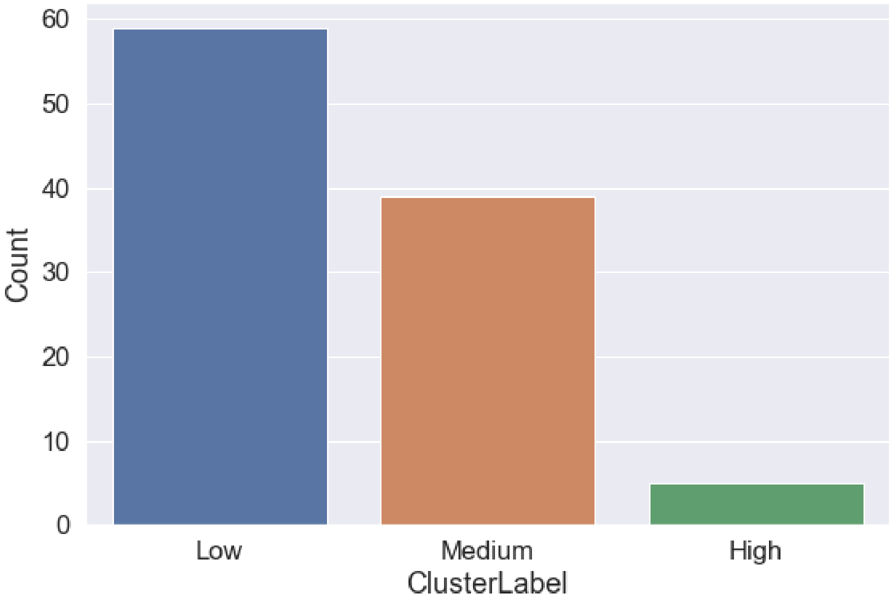
Cluster 2 has a **Medium** availability of venues:

Runnymede , Swansea	2	43.651571	-79.484450	44
Dufferin , Dovercourt Village	2	43.669005	-79.442259	44
Central Bay Street	2	43.657952	-79.387383	44
Harbourfront East , Union Station , Toronto Is...	2	43.640816	-79.381752	46
Queen's Park , Ontario Provincial Government	2	43.662301	-79.389494	48
Commerce Court , Victoria Hotel	2	43.648198	-79.379817	48
Parkdale , Roncesvalles	2	43.648960	-79.456325	48
Church and Wellesley	2	43.665860	-79.383160	49
Regent Park , Harbourfront	2	43.654260	-79.360636	50
Garden District, Ryerson	2	43.657162	-79.378937	50

Cluster 3 has a **High** availability of venues:

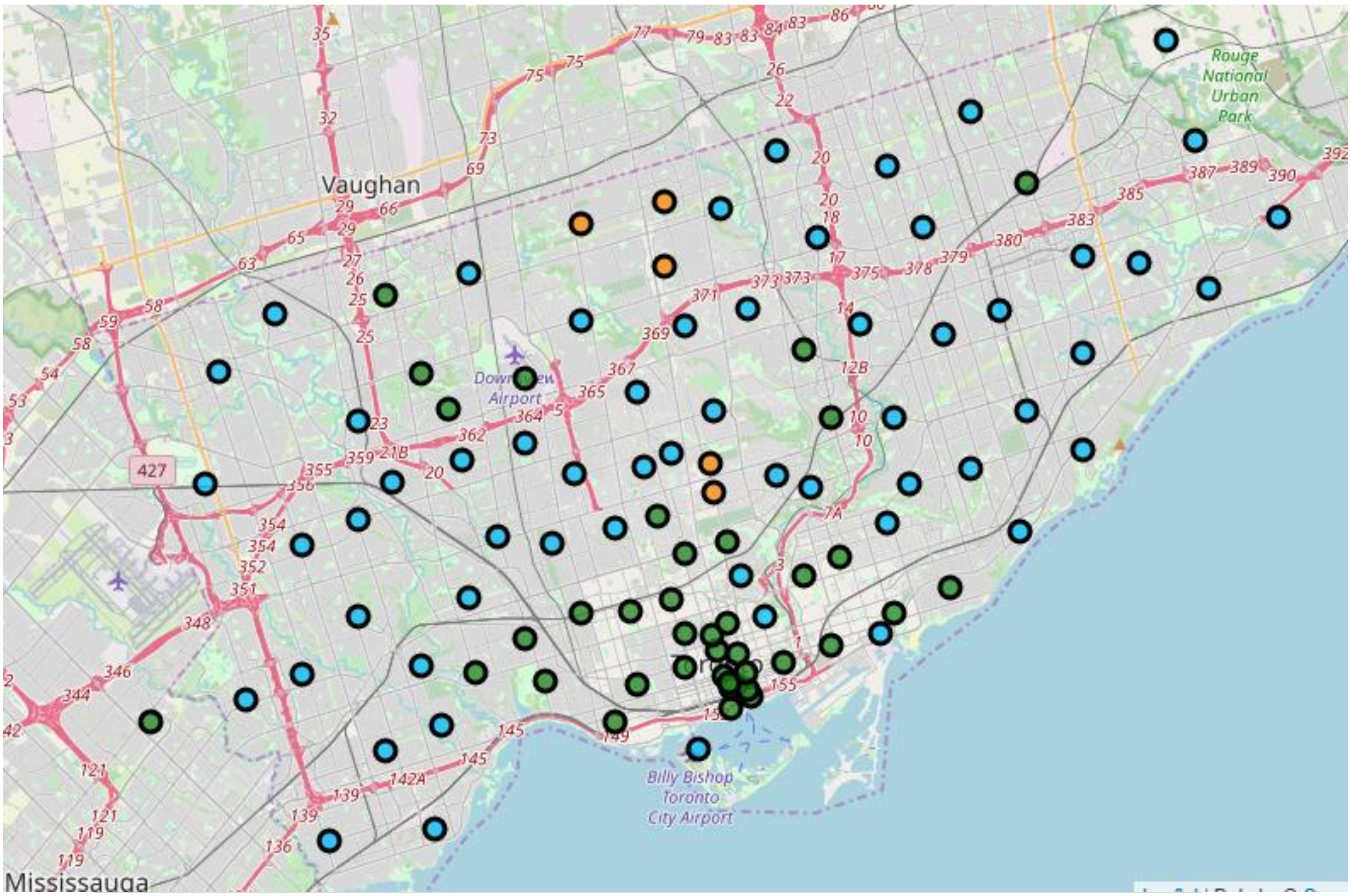
Willowdale	1	43.782736	-79.442259	95
Willowdale , Newtonbrook	1	43.789053	-79.408493	95
Davisville	1	43.704324	-79.388790	117
Davisville North	1	43.712751	-79.390197	117

We can visualize the number of each Clusters in a bar graph:



Cluster	Label	Color
0	Low Availability	Blue
2	Medium Availability	Orange
1	High Availability	Green

Finally, we can visualize the clustered Neighborhoods formed on the Map of Toronto:



5. Discussion

The aim of this project is to provide a solution leveraging the principles of Data Science to help people identify the neighborhoods to which they want to start a business or relocate to, depending on the availability of the most essential venues.

Depending on the choice of Features (venues in this case), the Clustering can be modified to accommodate the needs. The venues can be added or removed according to what the Use case is.

In the current scenario, if a person wishes to live in a quiet neighborhood with less footfall, Cluster 0 (Low) could be the ideal place. Similarly, a new Restaurant can be opened in any of the Cluster 0 neighborhoods as there is a scarcity of Restaurants in this zone. If one wants to relocate to a more happening neighborhood, Cluster 1(High) would be his choice.

6.Conclusion

This project aims to provide a solution to a User to get a better analysis of the neighborhoods with respect to the availability of venues. In future, this idea can be extended to many domains which leverage geographical data. This could be a Cab service, wishing to identify regions with high demand or a Tourism company who could provide better service by identifying high rated spots and tourist attraction zones in a City. The possibilities are endless.

With more data and more hours dedicated to this project, it can be developed as a Web or Mobile Application while keeping the basic idea similar to what was described and illustrated in the Report and the associated Notebook.