Capstone Project - The Battle of Neighbourhoods

Data Analysis of Geo-Location data to start-up a Cafe in Toronto City

-Tejas Mulay

1. Introduction:

1.1 Background

In the competitive world, it is of utmost importance to know the surroundings and take necessary steps leading to the success of a venture. A similar case exists among the food industry where a large number of start-ups emerge and may succeed or fail to depend on a lot of factors.

Such is the case with Cafés, which are currently popular among youths and profitable business for entrepreneurs. Yet, a decent location and favorable surroundings contribute heavily to the success.

1.2 Problem

Data acquisition of geo-location using foursquare and boroughs data o Wikipedia page to analyze and cluster the existing cafés and acquiring favorable locations to start-up a café.

This project aims to analyze the existing geo-locations of the cafés in Toronto city and clustering the cafés using tools like Python, Jupyter Notebook, Foursquare data, and Machine learning algorithms.

1.3 Interest

The entrepreneurs who are willing to venture into a café at Toronto City in exploring the geo-locations of existing cafés and acquiring a profitable location.

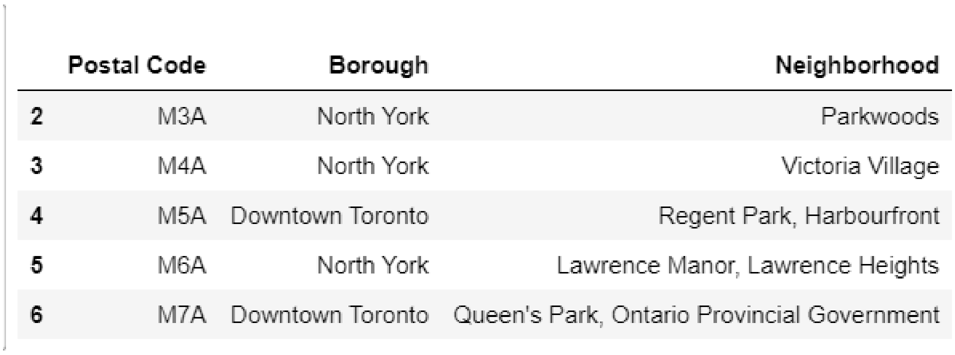
2. Data Acquisition and Cleaning

2.1 Data Acquisition

The data acquired for this project is a combination of two sources. The first data source being Wikipedia data on boroughs of Ontario and second, being foursquare data.
Wikipedia link to data: Link
Foursquare: Link

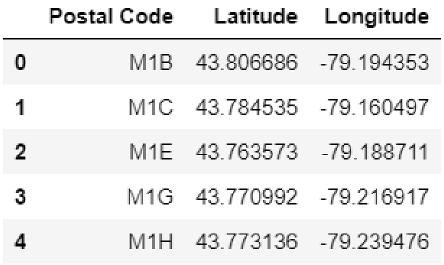
2.2 Data Cleaning:

The dataset of boroughs of Ontario can be scraped from Wikipedia page using pandas or beautiful soap library in python. The following is the screenshot of the dataset on boroughs.



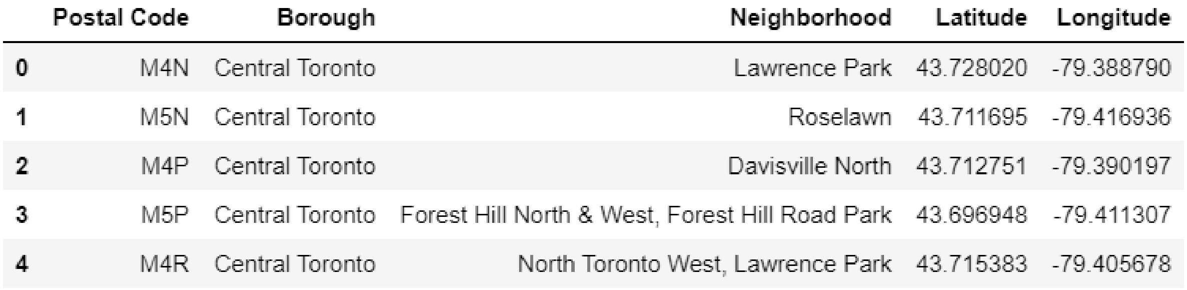
	Postal Code	Borough	Neighborhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Later the geocoder data of longitude and latitudes are acquired for the given location of postal code.



	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

Merging the datasets to get the location data of the specific borough, results in the following dataset-



	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790
1	M5N	Central Toronto	Roselawn	43.711695	-79.416936
2	M4P	Central Toronto	Davisville North	43.712751	-79.390197
3	M5P	Central Toronto	Forest Hill North & West, Forest Hill Road Park	43.696948	-79.411307
4	M4R	Central Toronto	North Toronto West, Lawrence Park	43.715383	-79.405678

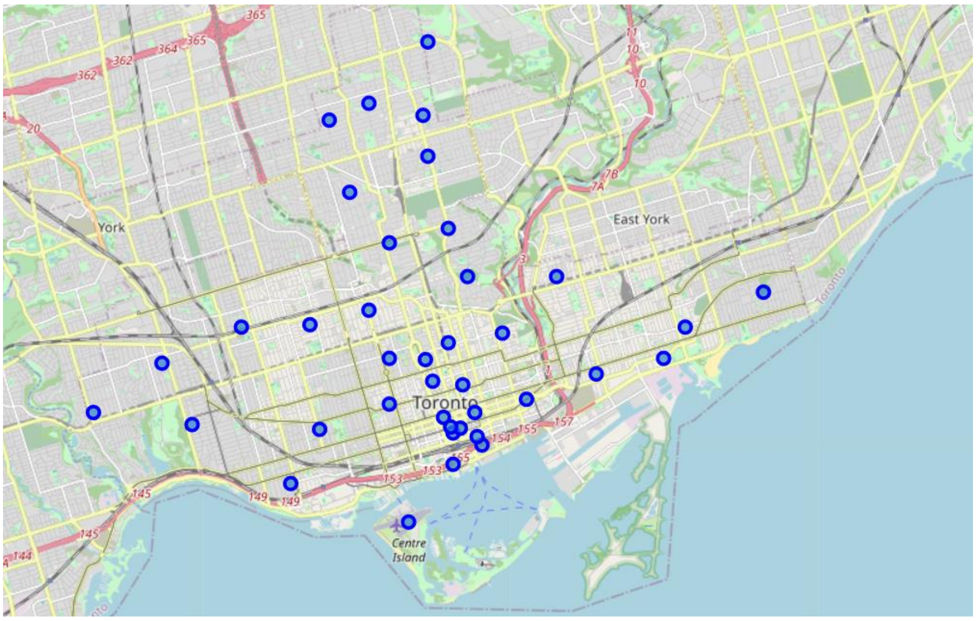
The data is later on used to acquire the foursquare data on venues around the region. Thus further leading to the clustering of data to analyze the region and acquiring potential geo-location.

3. Methodology:

3.1 Exploratory Data Analysis:

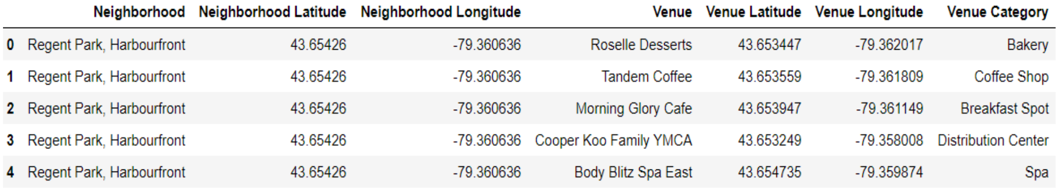
3.1.1 The neighborhood of Toronto:

The neighborhoods are plotted on the map using folium maps and the location of each neighborhood is observed as well as validated to get an idea of clustering to be performed further.



3.1.2 Boroughs of Toronto:

After acquiring foursquare data on venues around Toronto City the data was analyzed to get the category the neighborhood belongs to.

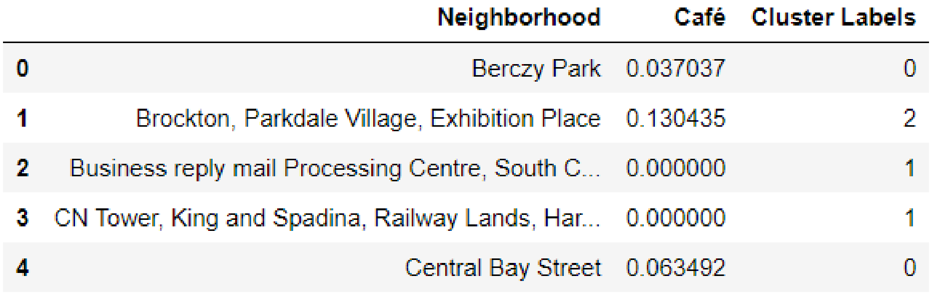


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Morning Glory Cafe	43.653947	-79.361149	Breakfast Spot
3	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
4	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa

The categories of all the neighborhoods being attained are observed by getting a count and the category 'Café' was selected for analysis.

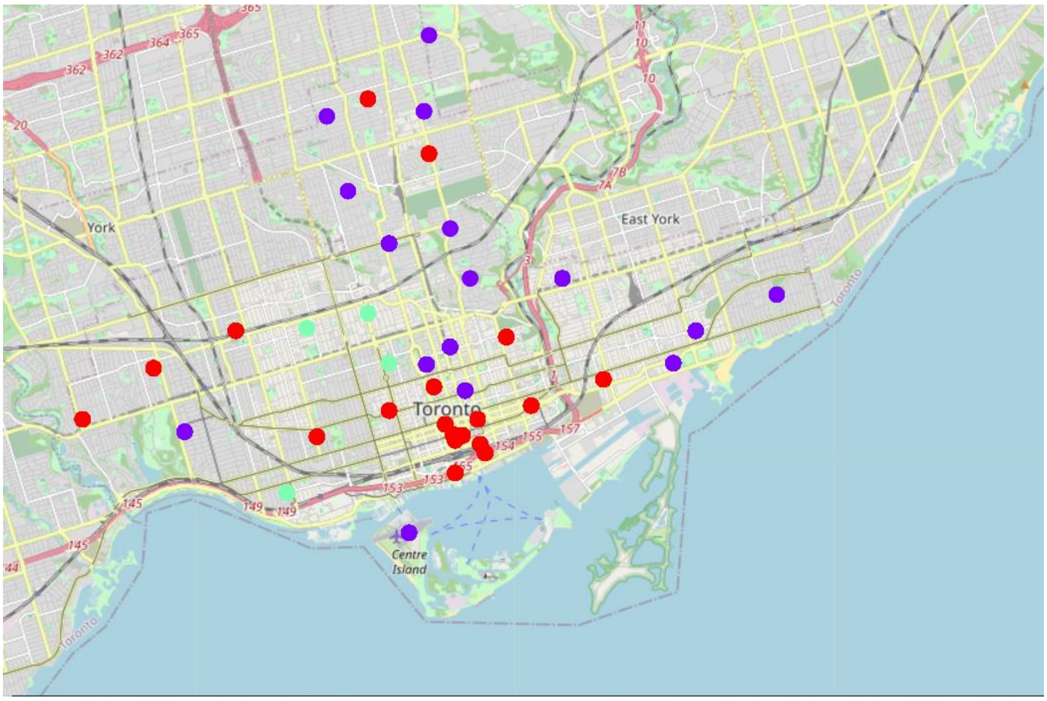
3.2 Clustering:

The data is clustered for the neighborhoods with category 'Café' using the k-means algorithm with k=3. The clustered data was then labeled according to their cluster and merged up with the location data.



	Neighborhood	Café	Cluster Labels
0	Berczy Park	0.037037	0
1	Brockton, Parkdale Village, Exhibition Place	0.130435	2
2	Business reply mail Processing Centre, South C...	0.000000	1
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	1
4	Central Bay Street	0.063492	0

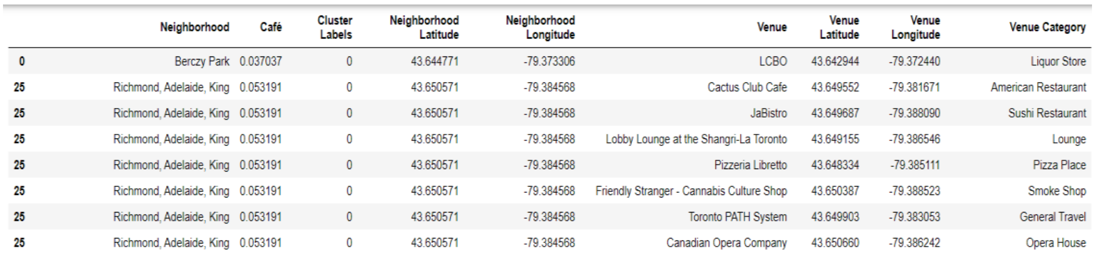
Thus formed table with the foursquare location data and geo-location data was then used to plot on the map using folium maps with an independent color to each cluster.



4. Results:

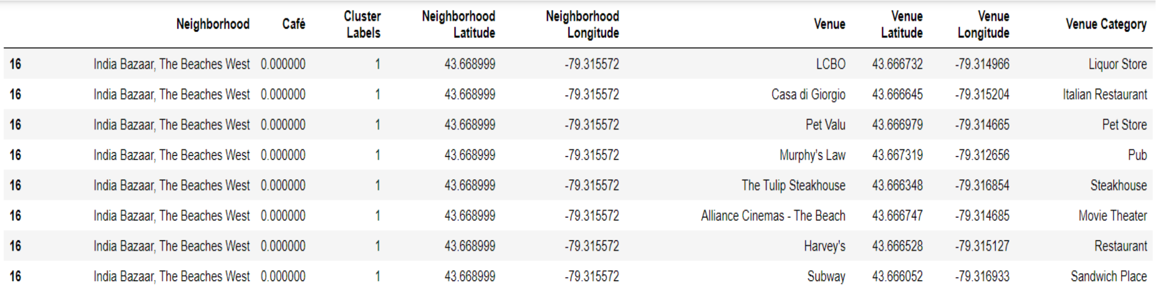
After running the K-means clustering the data is clustered into 3 clusters which can be looked upon by observing the map or the datasets-

Cluster-0



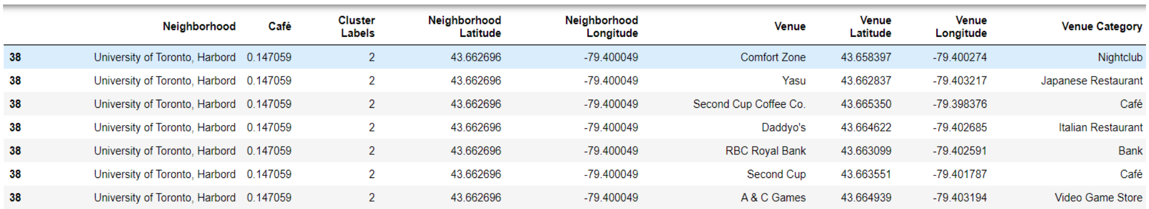
	Neighborhood	Café	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Berczy Park	0.037037	0	43.644771	-79.373306	LCBO	43.642944	-79.372440	Liquor Store
25	Richmond, Adelaide, King	0.053191	0	43.650571	-79.384568	Cactus Club Cafe	43.649552	-79.381671	American Restaurant
25	Richmond, Adelaide, King	0.053191	0	43.650571	-79.384568	JaBistro	43.649687	-79.388090	Sushi Restaurant
25	Richmond, Adelaide, King	0.053191	0	43.650571	-79.384568	Lobby Lounge at the Shangri-La Toronto	43.649155	-79.386546	Lounge
25	Richmond, Adelaide, King	0.053191	0	43.650571	-79.384568	Pizzeria Libretto	43.648334	-79.385111	Pizza Place
25	Richmond, Adelaide, King	0.053191	0	43.650571	-79.384568	Friendly Stranger - Cannabis Culture Shop	43.650387	-79.388523	Smoke Shop
25	Richmond, Adelaide, King	0.053191	0	43.650571	-79.384568	Toronto PATH System	43.649903	-79.383053	General Travel
25	Richmond, Adelaide, King	0.053191	0	43.650571	-79.384568	Canadian Opera Company	43.650660	-79.386242	Opera House

Cluster-1



	Neighborhood	Café	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
16	India Bazaar, The Beaches West	0.000000	1	43.668999	-79.315572	LCBO	43.666732	-79.314966	Liquor Store
16	India Bazaar, The Beaches West	0.000000	1	43.668999	-79.315572	Casa di Giorgio	43.666645	-79.315204	Italian Restaurant
16	India Bazaar, The Beaches West	0.000000	1	43.668999	-79.315572	Pet Valu	43.666979	-79.314665	Pet Store
16	India Bazaar, The Beaches West	0.000000	1	43.668999	-79.315572	Murphy's Law	43.667319	-79.312656	Pub
16	India Bazaar, The Beaches West	0.000000	1	43.668999	-79.315572	The Tulip Steakhouse	43.666348	-79.316854	Steakhouse
16	India Bazaar, The Beaches West	0.000000	1	43.668999	-79.315572	Alliance Cinemas - The Beach	43.666747	-79.314685	Movie Theater
16	India Bazaar, The Beaches West	0.000000	1	43.668999	-79.315572	Harvey's	43.666528	-79.315127	Restaurant
16	India Bazaar, The Beaches West	0.000000	1	43.668999	-79.315572	Subway	43.666052	-79.316933	Sandwich Place

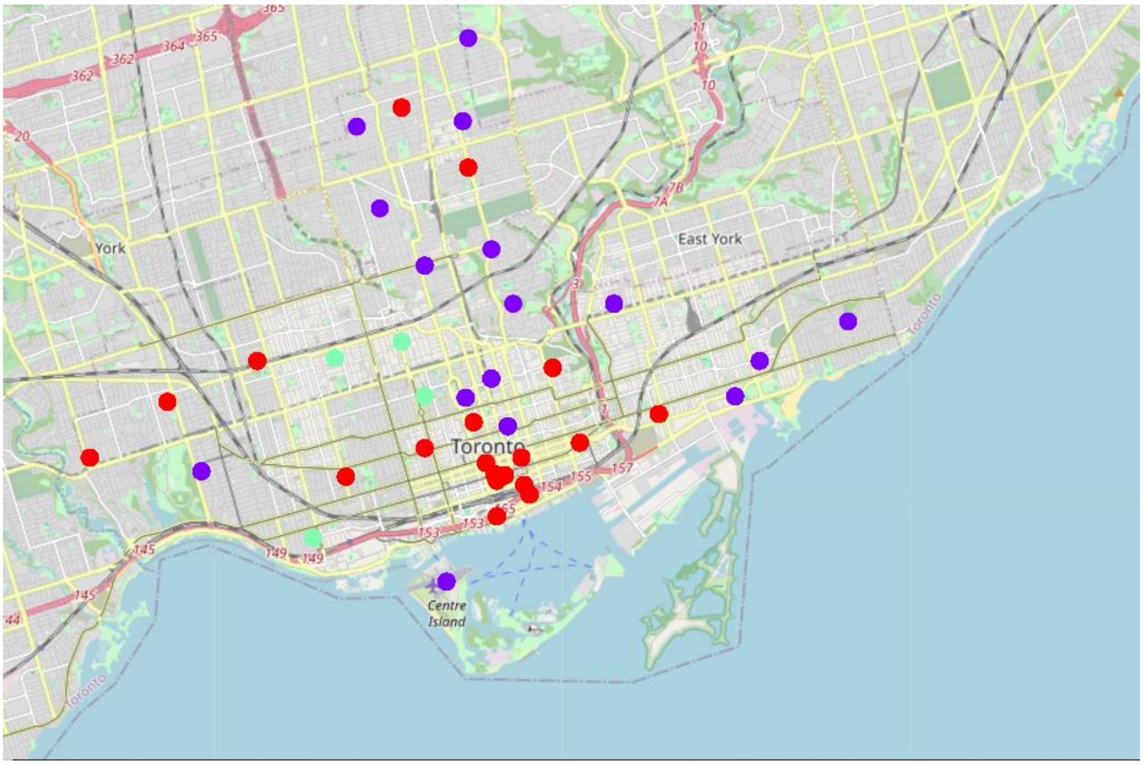
Cluster-2



	Neighborhood	Café	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
38	University of Toronto, Harbord	0.147059	2	43.662696	-79.400049	Comfort Zone	43.658397	-79.400274	Nightclub
38	University of Toronto, Harbord	0.147059	2	43.662696	-79.400049	Yasu	43.662837	-79.403217	Japanese Restaurant
38	University of Toronto, Harbord	0.147059	2	43.662696	-79.400049	Second Cup Coffee Co.	43.665350	-79.398376	Café
38	University of Toronto, Harbord	0.147059	2	43.662696	-79.400049	Daddyo's	43.664622	-79.402685	Italian Restaurant
38	University of Toronto, Harbord	0.147059	2	43.662696	-79.400049	RBC Royal Bank	43.663099	-79.402591	Bank
38	University of Toronto, Harbord	0.147059	2	43.662696	-79.400049	Second Cup	43.663551	-79.401787	Café
38	University of Toronto, Harbord	0.147059	2	43.662696	-79.400049	A & C Games	43.664939	-79.403194	Video Game Store

Map:

Cluster-0: Red - Neighbourhoods with more Cafés
Cluster-1: Purple - Neighbourhoods with more Cafés
Cluster-2: Light Green - Neighbourhoods with very few Cafés



5. Discussion:

Based on the café geolocation and venue foursquare data, the visualization of clustered data can be done using folium maps. The following insights can be drawn from the clustered data. The cluster-0 and cluster-1 seem way denser I comparison to cluster-2. As a result, there are more no of existing cafés in those region leading to the recommendation of the cluster-2 neighborhood to be utilized for starting a new café. The cluster-2 region includes the following neighborhood- University of Toronto, Harbord, Christie, Brockton, Parkdale Village, Exhibition Place.

6. Conclusion:

The project enables entrepreneurs to get a better understanding of the neighborhood for the existing Cafés. This helps them to gain knowledge of the surroundings and to chalk out a region or neighborhood to venture into a new enterprise in the form of a café. This project takes the help of technology and data to draw insights out of data to be one step ahead in the competition.