LOCATION RECOMMENDATION FOR A NEW BUSINESS UNIT

THE BATTLE OF NEIGHBORHOODS

DATA SCIENCE CAPSTONE PROJECT

SUBMITTED BY

BHAVANI MADDULA

Business Problem

Statement

Toronto is the capital of the province of Ontario and is one of the most populous cities in Canada. Apart from being the financial capital of the country, Toronto is home for arts and culture, and is thus recognised as one of the most multicultural and cosmopolitan cities in the world. Toronto is one of the most sought-after destination for immigrants from several years making it rich in its ethnic diversity.

Over the past two decades, immigration numbers increased significantly, with the maximum number of immigrants coming from South Asian origin. Among these countries, Indians rank first in the list and Indo-Canadians are one of the fastest growing communities in Toronto.

Owing to the rich cultural and diverse heritage of Toronto, a Business entrepreneur wants to open an 'Indian Arts, Dance and Cultural Centre' in one of the neighborhoods. This centre would adhere to the needs of residents with Indian origin by providing facilities for learning different art and dance forms. The centre would also organise several cultural events reflecting the vast and diverse heritage of India. The goal of this project would be to help this entrepreneur identify a suitable location for establishing this business unit.

Solution to the Business Problem

It is a well-known fact that the success of any business depends on a broad spectrum of factors. Apart from the services offered by the company, the entrepreneur should focus on the location, demographics of a given area, neighborhoods and other specific factors related to the business domain. A proper location can then be identified by analyzing all the related aspects. This is a very crucial step that would ensure success for the business at hand.

The ideal location for establishing the business unit under discussion would be in a neighborhood that is mostly inhabited by people with Indian origin or closer to such neighborhoods. This location should be well connected, and restaurants and coffee shops should be in the vicinity, so that food needs of the customers could be easily met. To get the location of restaurants, coffee shops and other amenities in each neighborhood, Foursquare location data can be used.

Stakeholders:

The stakeholders for this business solution would be any entrepreneur or government body interested in establishing new business units or centers catering to the needs of a given ethnic community.

Data Section

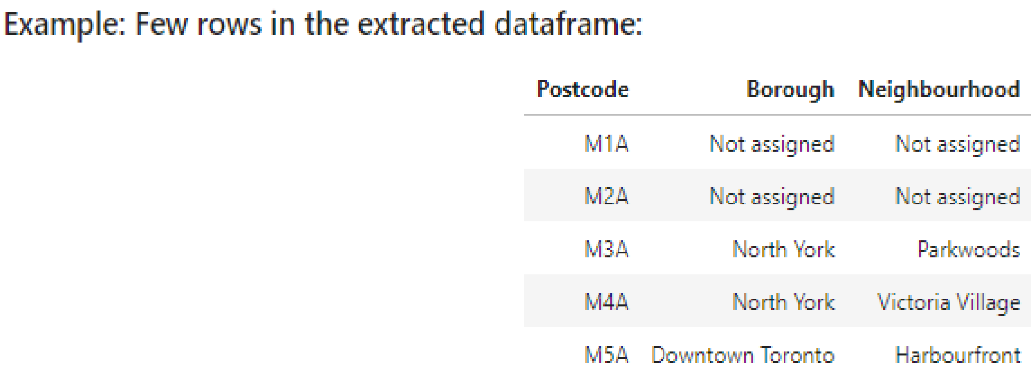
Data Sources

To address the business problem discussed, data sets will be generated from the following sources:

1. Toronto neighborhoods
   https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
2. Demographics of Toronto neighborhoods
   https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighborhoods
3. Geocoder/Google geolocation API/Geo spatial coordinates csv file
4. Foursquare APIs

Data Sources at a glance:

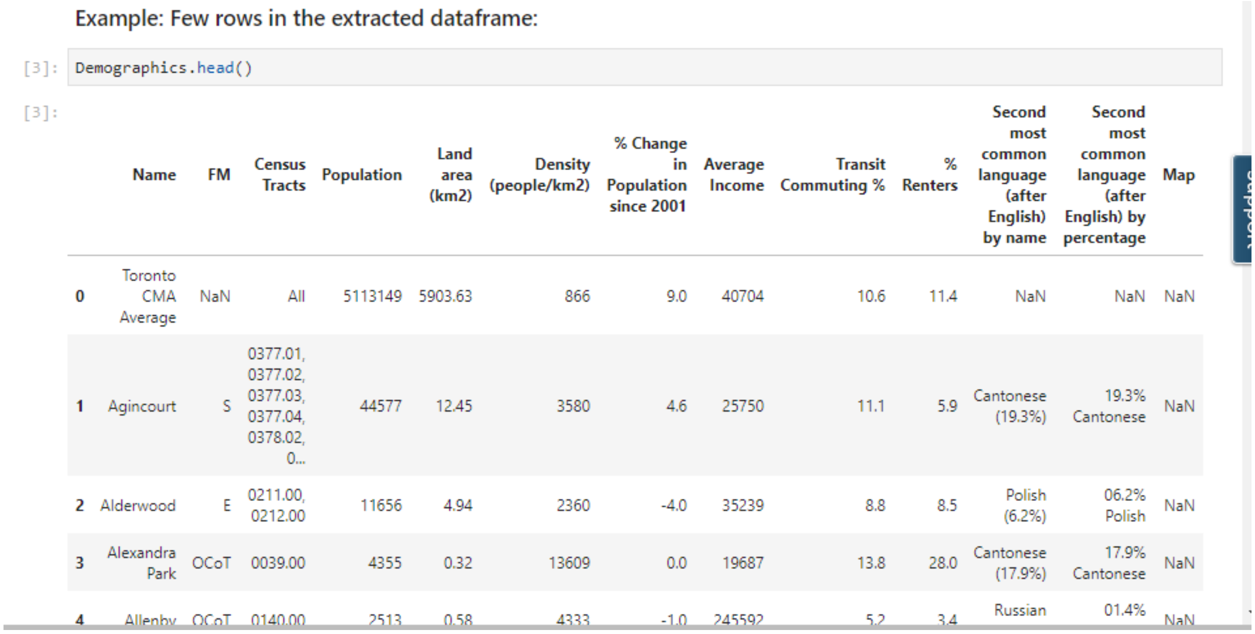
1. **Toronto neighborhoods**: This Wikipedia page provides a list of postal codes of Canada beginning with the letter M. All these postal codes correspond to Boroughs and neighborhoods located within Toronto. The required data set is extracted using ‘Beautiful soup’ and this is one of the data sets that will be used to solve the problem at hand.



Example: Few rows in the extracted dataframe:

Postcode	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Harbourfront

2. **Demographics of Toronto neighborhoods**: Apart from the data source listed above, demographic information is also quite crucial for addressing the given problem as the business venture is primarily focussed on an ethnic group. This Wikipedia page provides a list of demographic information for Toronto neighborhoods. ‘Beautiful soup’ will be used to extract the required data and then generate the required data set.



Example: Few rows in the extracted dataframe:

```
[3]: Demographics.head()
```

[3]:

	Name	FM	Census Tracts	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	% Renters	Second most common language (after English) by name	Second most common language (after English) by percentage	Map
0	Toronto CMA Average	NaN	All	5113149	5903.63	866	9.0	40704	10.6	11.4	NaN	NaN	NaN
1	Agincourt	S	0377.01, 0377.02, 0377.03, 0377.04, 0378.02, 0...	44577	12.45	3580	4.6	25750	11.1	5.9	Cantonese (19.3%)	19.3% Cantonese	NaN
2	Alderwood	E	0211.00, 0212.00	11656	4.94	2360	-4.0	35239	8.8	8.5	Polish (6.2%)	06.2% Polish	NaN
3	Alexandra Park	OCoT	0039.00	4355	0.32	13609	0.0	19687	13.8	28.0	Cantonese (17.9%)	17.9% Cantonese	NaN
4	Allenby	OCoT	0140.00	2513	0.58	4333	-1.0	245592	5.2	3.4	Russian	01.4%	NaN

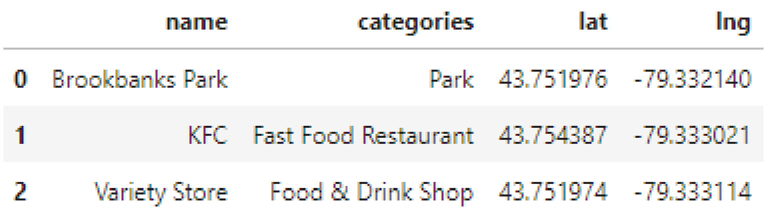
3. **Geocoder/ Google geolocation API**: Address geocoding refers to the process of finding an associated latitude and longitude for a given address. The geocoordinates (viz. latitude and longitude) for the neighborhoods will be obtained using the geocoder or by using Google geolocation API or from the geospatial coordinates.csv file.

   Example: The following information is returned for the coordinates of Downtown Toronto The geographical coordinates of Downtown Toronto are 43.6541737, -79.3808116451341.

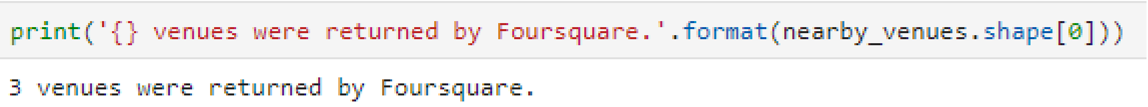
4. **Foursquare API**: Foursquare is one of the most popular Location Based Social Network (LBSN) in recent times. Foursquare provides personalized recommendations of places to go to near a user's current location based on users' previous browsing history, purchases, or check-in history. It allows users to explore the world around them and provides geo

tagged information. The Foursquare API allows application developers to interact with the Foursquare platform. The API provides location-based experiences with diverse information about venues, users, photos, and check-ins. The API supports real time access to places, Snap-to-Place that assigns users to specific locations, and Geo-tag. API calls will be made to obtain the required information for different venues of interest located in the neighborhoods. This information is crucial in meeting the objective of this project.

Example: The following result was returned for the venues in 'Parkwoods' neighborhood.



	name	categories	lat	lng
0	Brookbanks Park	Park	43.751976	-79.332140
1	KFC	Fast Food Restaurant	43.754387	-79.333021
2	Variety Store	Food & Drink Shop	43.751974	-79.333114



```
print('{} venues were returned by Foursquare.'.format(nearby_venues.shape[0]))
```

```
3 venues were returned by Foursquare.
```

Methodology

The dataframe corresponding to demographics of Toronto has 175 rows and 13 columns. After performing data cleaning, exploratory data analysis was carried out to gain insights into the neighborhood's population, population density corresponding to neighborhoods with different languages.

Note: Language here represents the second most popular language after English in a given neighborhood.

The dataframe corresponding to Neighborhoods of Toronto has 289 rows and 3 columns, and the Geospatial coordinates.csv file contains 103 rows. This csv file contains geospatial information for different postal codes in Toronto. Geocoder couldn't be used in this project, as it was issuing multiple warnings, with no desirable results. After performing data cleaning, the three dataframes

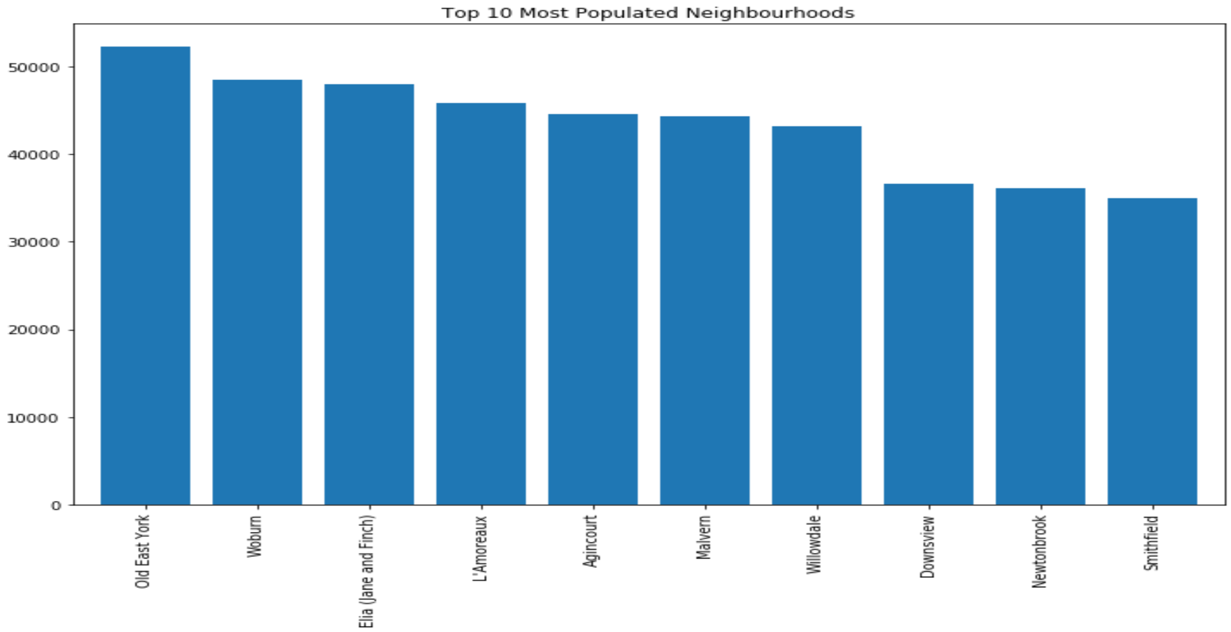
(Demographics of Toronto, Neighborhoods of Toronto, Geospatial coordinates for neighborhoods of Toronto) were merged to get the final dataframe for further analysis.

Using Foursquare credentials, the top 100 venues in each neighborhood were obtained and 222 unique categories were identified. Next, a dataframe listing the top 10 venues for each neighborhood was created. This dataframe serves as the data set for performing clustering at the stage.

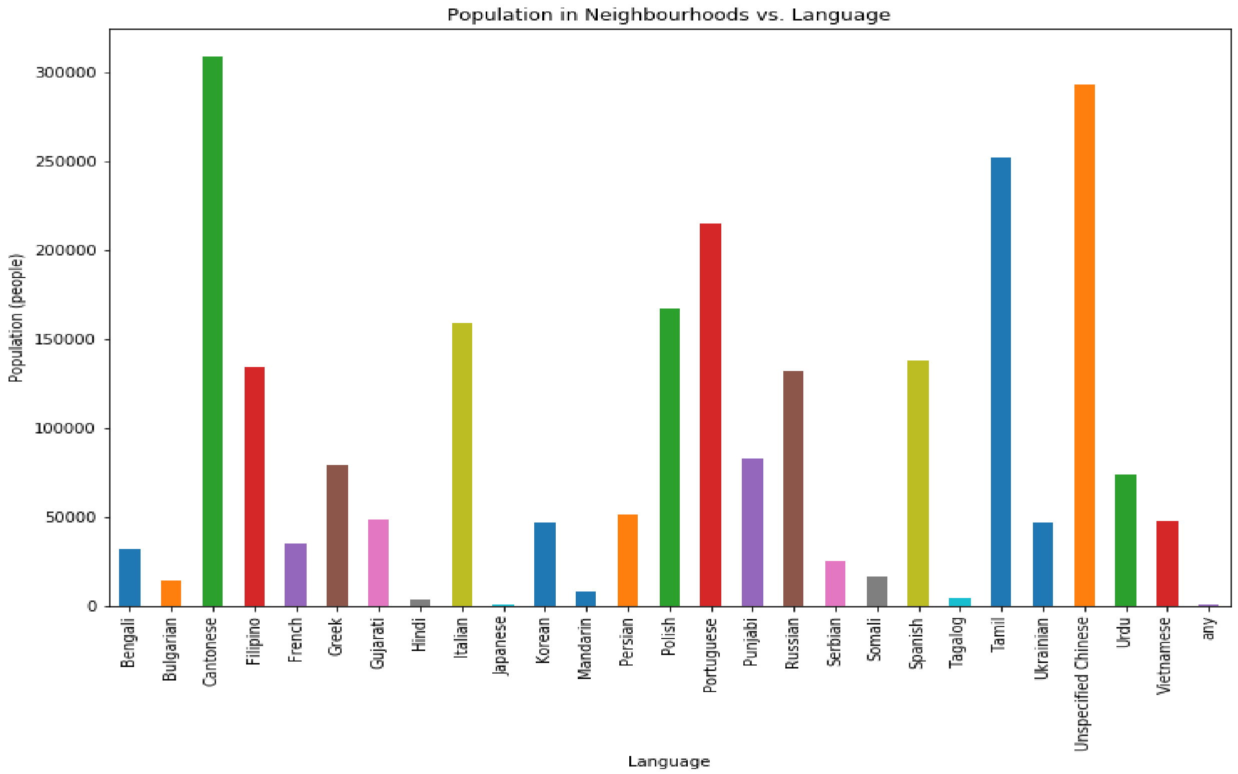
Exploratory Data Analysis

1. Top 10 most populated neighborhoods

Here is the plot for the top 10 most populated Neighborhoods in the city of Toronto.



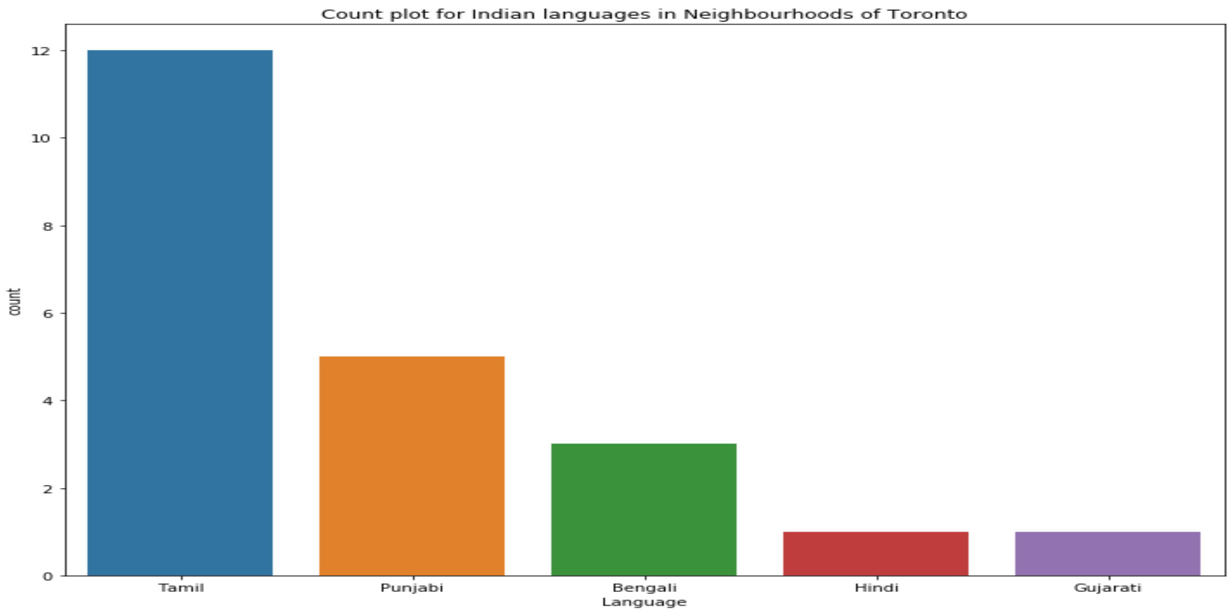
2. **Neighborhood population versus the second most popular language (after English) in the neighborhood**



The above plot shows that 'Portuguese' is the most popular language and out of all the Indian languages, 'Tamil' is the most popular language.

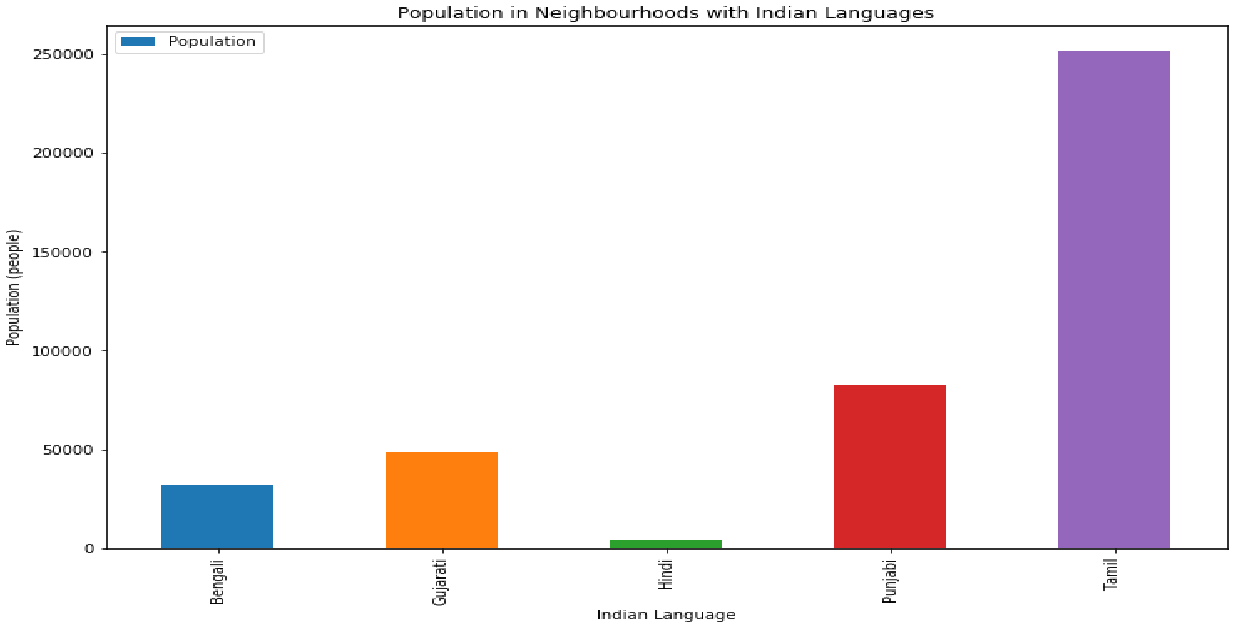
Note: Bengali, Punjabi, Gujarati, Hindi, Tamil are Indian Languages

3. Indian languages in different neighborhoods of Toronto



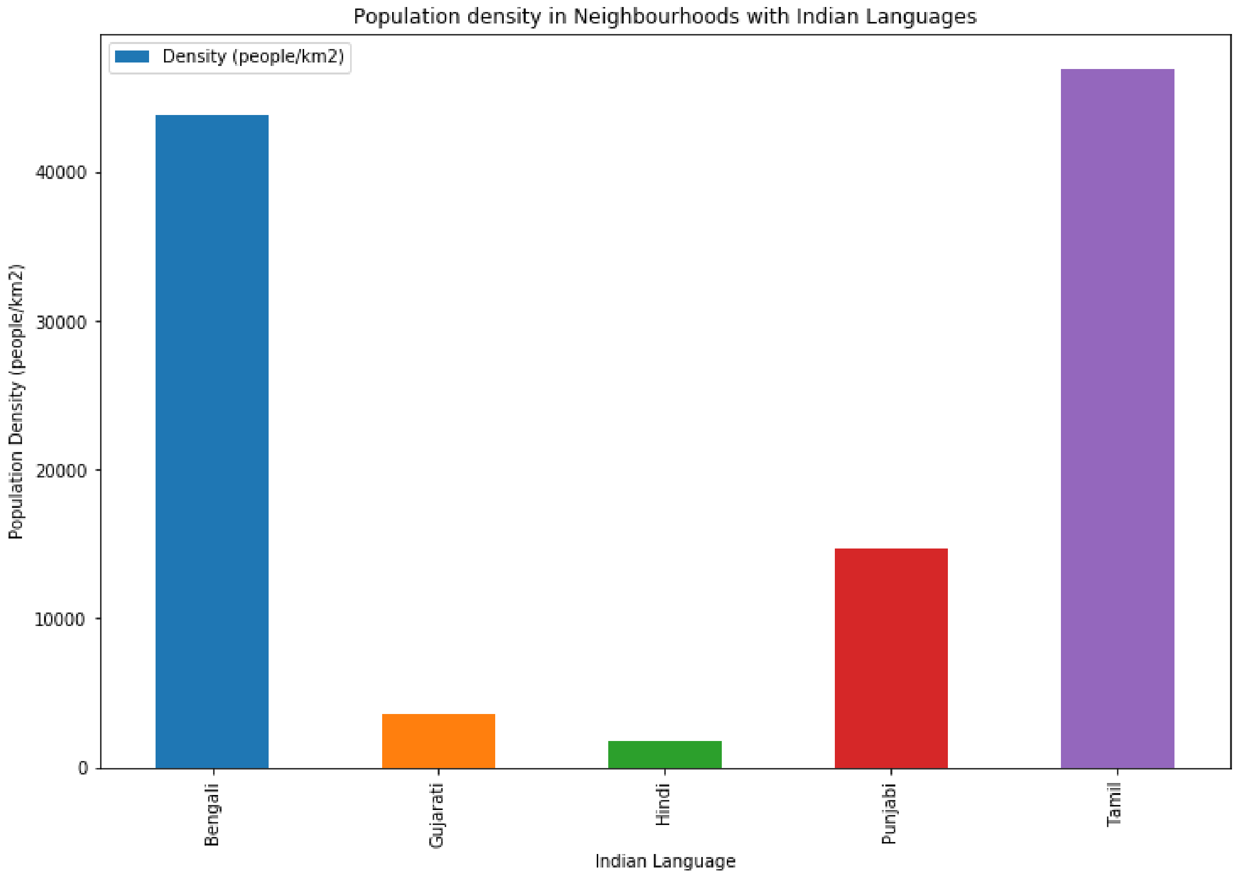
The above plot clearly shows that among the Indian languages, Tamil language is the most popular one among the neighborhoods.

4. Population in neighborhoods with Indian languages



From the above plot, we can conclude that most of the Indian population in Toronto speak Tamil.

5. Population density in neighborhoods with Indian languages

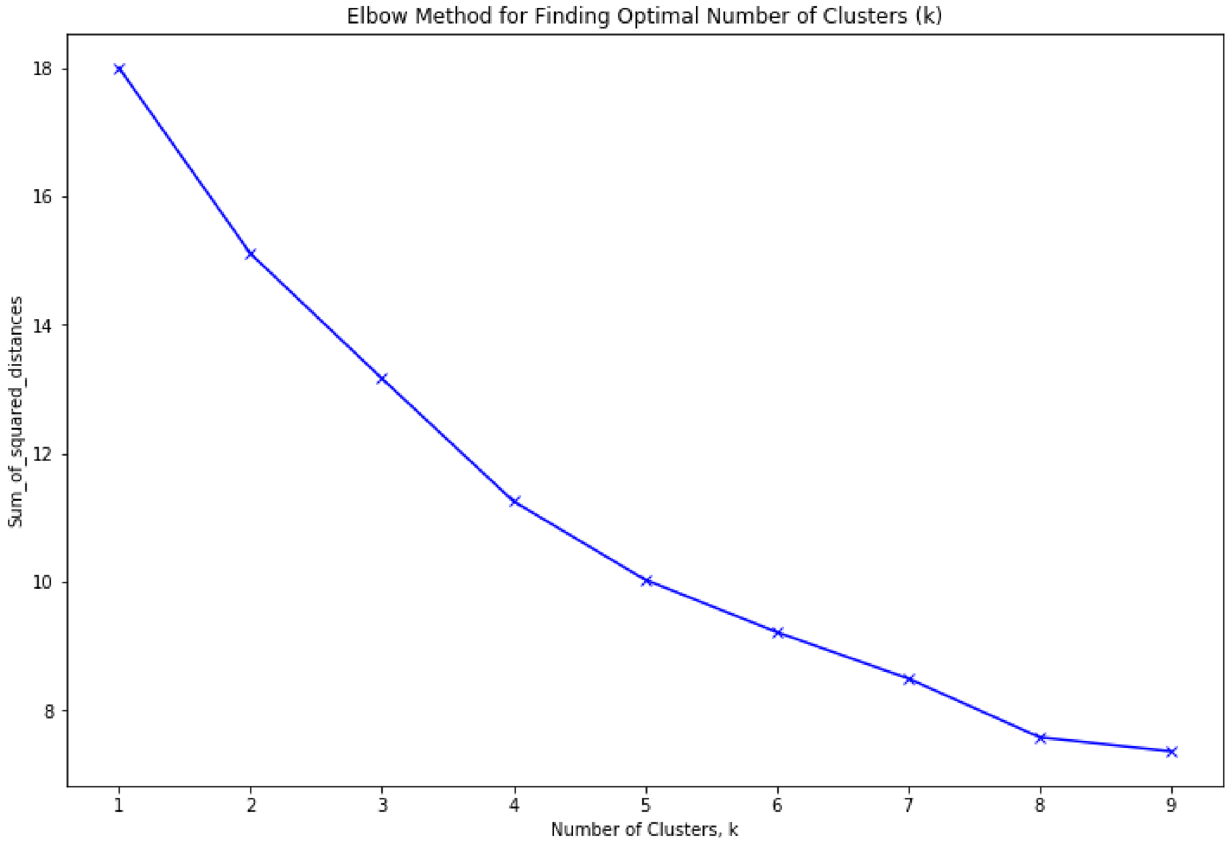


The above plot shows that Tamil neighborhoods in the city of Toronto are densely populated.

Inference from EDA:

From the above plots, we can infer that identifying a neighborhood where Tamil is the most popular language (after English) and with good number of restaurants would be a good location for establishing our business unit.

K means Algorithm: The goal of this study is to identify a suitable neighborhood among several neighborhoods. This is an unsupervised machine learning problem and clustering techniques can be used to solve the same. In this project, K means clustering was used for clustering the data, because k means is a simple, inexpensive and efficient when working with large data sets. Further, the ‘elbow method’ was used to find the optimum value for the number of clusters, ‘k’. Clustering was then performed on the data set to identify the neighborhoods meeting the requirements addressed in the business problem, i.e. neighborhoods with Indian language and with restaurants in the vicinity.

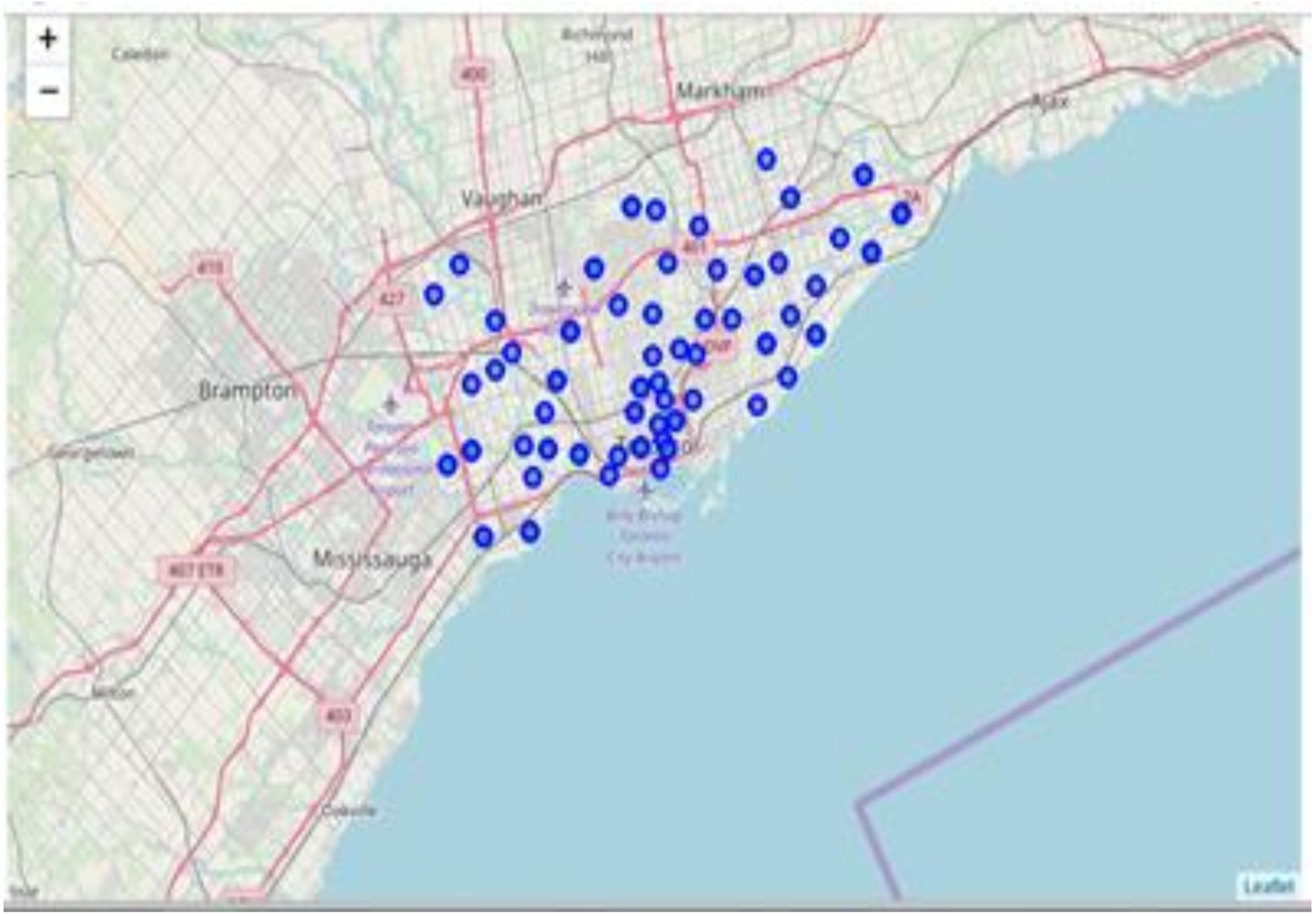


From the above graph, it is evident that the optimum number of clusters is 5.

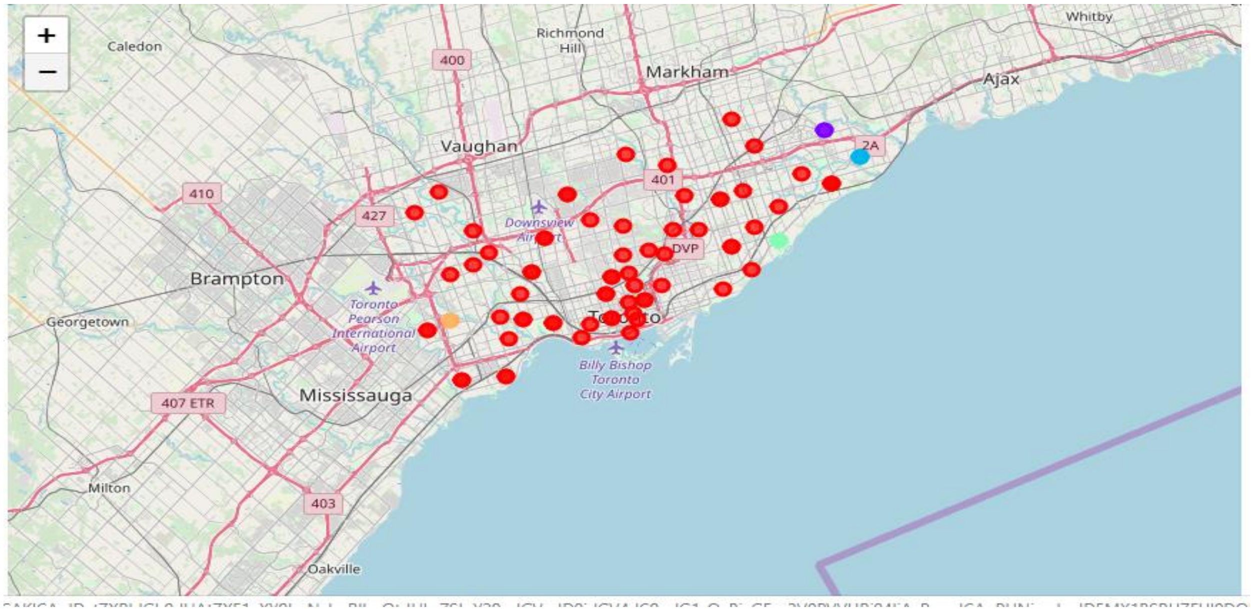
Results

Each cluster was analysed to identify neighborhoods with Indian language and restaurants in the vicinity. Few neighborhoods meeting the language requirement were observed. However, these neighborhoods do not include many restaurants. The result shows that **'Dorset Park'** is a neighborhood with Tamil language and relatively includes greater number of restaurants. The business statement requires a location with restaurants in the vicinity. So, this neighborhood can be considered for opening the new business unit.

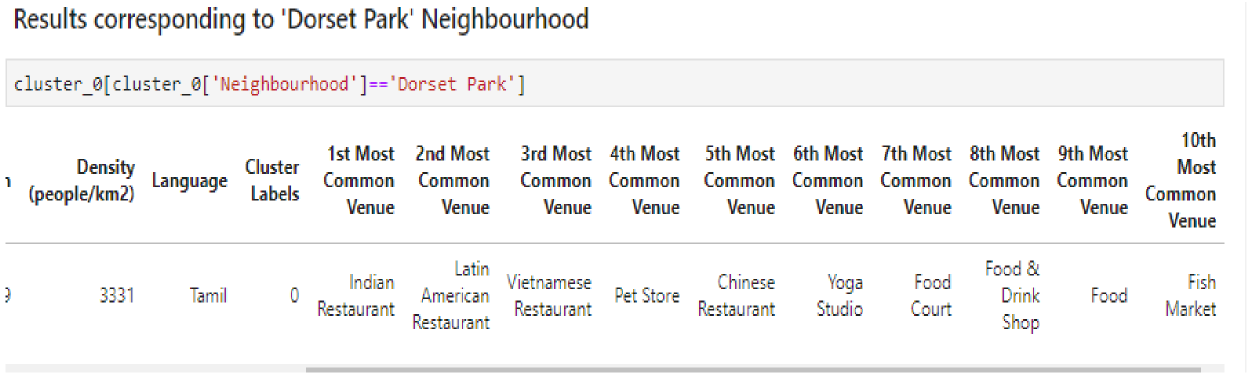
Neighborhoods of Toronto



Clusters of Neighbourhoods (k = 5)



10 Most Common Venues in 'Dorset Park' Neighborhood



Results corresponding to 'Dorset Park' Neighbourhood

```
cluster_0[cluster_0['Neighbourhood']=='Dorset Park']
```

ı	Density (people/km2)	Language	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
9	3331	Tamil	0	Indian Restaurant	Latin American Restaurant	Vietnamese Restaurant	Pet Store	Chinese Restaurant	Yoga Studio	Food Court	Food & Drink Shop	Food	Fish Market

Discussion

Among the 5 clusters, the first cluster is very dense and includes 67 neighborhoods. These neighborhoods include several restaurants, pizza points and coffee shops. The second cluster includes 2 identical neighborhoods and the third cluster comprises of 3 similar neighborhoods. 2 neighborhoods each form the fourth and fifth clusters.

In this project, the geospatial coordinates corresponding to all the neighborhoods listed in Demographics of Toronto dataframe couldn't be obtained and this resulted in loss of few neighborhoods in the merged dataframe. However, if all the neighborhoods could be retained, a much richer data set can be obtained which would further result in a better analysis.

Conclusion

Data visualization turns out to be powerful in drawing insights from data and towards addressing the business problem at hand. Location Based Social Networks like Foursquare provide the flexibility to solve several interesting problems that would in turn benefit the stakeholders. In this capstone project, exploratory data analysis and Foursquare API data were used to identify a neighborhood for opening an ‘Indian Arts, Dance and Cultural Centre’. Innovative business solutions can be obtained by performing similar analysis on many problems related to several fields.